Université de Montréal

Architecture question-réponse pour l'automatisation des services d'information

par
Luc Bélanger

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph. D.)
en informatique

Août, 2006

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :
Architecture question-réponse pour l'automatisation des services d'information

présentée par :
Luc Bélanger

a été évaluée par un jury composé des personnes suivantes :

Jian-Yun Nie
président-rapporteur

Guy Lapalme
directeur de recherche

Balázs Kégl
membre du jury

Marc El-Bèze, Université d'Avignon
[redacted]

acceptée le :

.....

Résumé

Cette thèse traite de l'automatisation des services d'information, et particulièrement du problème de la réponse automatisée aux courriels pour laquelle nous présentons une architecture inspirée de la question-réponse. Cette architecture a été développée dans le cadre du service de relations avec les investisseurs d'une grande entreprise publique. Notre solution à ce problème améliorera la qualité des services parce que la majorité des courriels seront traités automatiquement permettant ainsi aux autres de l'être avec plus d'attention.

Notre architecture est divisée en modules correspondant aux étapes de traitement d'un système de question-réponse factuelle : le prétraitement des documents de référence, le traitement des courriels, la sélection des sources de réponse, l'analyse des documents candidats et l'identification des réponses. La pertinence de l'architecture a été validée par des expériences réalisées sur certains modules de traitement.

Le lien entre le traitement automatisé des courriels et la question-réponse a été établi de deux manières : par une description d'un corpus de courriels et par l'analyse du procédé interrogatif de la question. Cette étude a montré que les questions contenues dans les courriels sont plus difficiles à traiter que les questions factuelles répondues par les systèmes de question-réponse.

Le développement d'un système basé sur l'architecture question-réponse exploite des approches diversifiées du traitement de la langue naturelle. Lors du traitement des courriels, nous considérons une approche symbolique à base de règles pour repérer les questions, tandis que lors de l'analyse des documents candidats, nous considérons une approche statistique pour repérer les rôles sémantiques d'un prédicat dans une phrase.

Les réponses aux questions contenues dans les courriels proviennent de diverses sources que nous coordonnons en décrivant une modélisation du domaine des relations avec les investisseurs et en référant au langage d'annotation TimeML pour annoter le temps et les événements. Nous décrivons ensuite la représentation graphique de ces annotations, utilisée dans un outil d'annotation de TimeML : TANGO.

Nous terminons la thèse en expliquant comment exploiter les composantes de l'architecture pour développer un système de réponse automatisée aux courriels et en décrivant les améliorations que nous pouvons apporter aux composantes de l'architecture.

Mots clés : question-réponse, réponse automatisée au courriel, traitement de la langue naturelle, étiquetage sémantique, représentation temporelle, système d'information

Abstract

This thesis deals with the automation of information services, especially with the problem of automated email answering for which we present an architecture based on question answering. This architecture was developed in the context of the investor relations services of a large public company. Our solution to this problem will improve the quality of those services, since it will allow most emails to be processed automatically, thus leaving more time to process the others.

Our architecture is divided into modules, each corresponding to a processing stage in a factual question answering system: preprocessing of reference documents, analysis of emails, selection of answer sources, analysis of candidate documents and identification of answer. The experiments we carried out on some modules validated the relevance of the architecture.

The link between automated email answering and question answering has been established in two ways: by a description of an email corpus and by an analysis of the interrogative focus of the question. This study showed that questions found in emails are harder to process than factual questions answered by other question answering systems.

The development of a system based on a question answering architecture benefits from different approaches to natural language processing. For email processing, we use a symbolic approach based on rules to identify the questions, while for candidate documents analysis, we rely on a statistical approach to label the semantic roles of a predicate within a sentence.

The answers to questions found in emails come from several sources that we coordinate by describing a model of the investor relations domain and refering to the annotation language TimeML to annotate time and events. We then describe a graphical representation of these annotations, which is used in the TANGO TimeML annotation tool.

We conclude by describing how we can use the components of our architecture to implement an automated email answering system and by suggesting some improvements.

Keywords: question answering, automated email answering, natural language processing, semantic role labeling, temporal representation, information system

Table des matières

Liste des tableaux

Table des figures

Remerciements

Je remercie mon directeur de recherche, Guy Lapalme, pour le soutien apporté lors de la réalisation de cette thèse. Je remercie également l'équipe du laboratoire de recherche en linguistique appliquée (RALI).

Je tiens aussi à remercier James Pustejovsky, Dragomir Radev, Inderjeet Mani, l'ensemble des participants aux ateliers de travail TERQAS et TANGO, de même que le Northeast Regional Research Center.

Je remercie les organismes suivants pour leur soutien financier lors de cette thèse : les Laboratoires Universitaires Bell, le Fonds québécois de la recherche sur la nature et les technologies, le Conseil de recherches en sciences naturelles et en génie du Canada, North Side Inc et Precarn Incorporated.

Chapitre 1

Introduction

L'automatisation des services d'information est essentielle pour les entreprises oeuvrant dans le domaine des services. Les usagers des services d'information sont exigeants, les ressources humaines sont limitées et la demande pour ces services ne cesse de croître. Auparavant, le moyen de communication privilégié pour avoir accès au service d'information était le téléphone. Grâce au courriel, il devient maintenant plus facile de contacter le service, ce qui augmente la quantité de requêtes à traiter. L'automatisation du service d'information rend le service plus efficace, en diminuant le temps de traitement d'une requête et en traitant ainsi un plus grand nombre de requêtes.

Dans cette thèse, nous portons notre attention sur une tâche parmi celles effectuées dans un service d'information : le traitement des courriels. Cette tâche propose plusieurs défis, et, d'un point de vue technique, elle peut être complètement automatisée. Il y a déjà sur le marché des systèmes de gestion des communications par courriels, mais ceux-ci ne permettent pas, globalement, de traiter automatiquement tous les courriels, de la réception à la rédaction de la réponse. Ces systèmes sont essentiels pour les services d'information, ils sont généralement utilisés pour traiter et archiver les courriels lorsque plusieurs personnes sont assignées au service. Les systèmes de gestion des courriels solutionnent principalement les problèmes techniques associés au traitement des courriels dans les services d'information. Ils facilitent le travail des préposés, mais ils pourraient être beaucoup plus efficaces en exploitant le fait que le courriel peut être analysé par des algorithmes et qu'il est possible

d'y répondre en consultant les données disponibles en format numérique. Les entreprises ne sont pas préparées pour le traitement complètement automatisé des courriels, la plupart d'entre elles n'ont même pas d'infrastructures pour archiver les communications électroniques.

La réponse automatisée aux courriels est essentielle pour les services d'information parce que le nombre de communications par courriel ne cesse d'augmenter. Présentement, un utilisateur corporatif moyen envoie et reçoit 133 courriels par jour ; ce nombre devrait atteindre 160 en 2009 [64]. Le courriel est un moyen de communication économique combinant les avantages de l'oral et de l'écrit, il est transmis instantanément et il laisse une trace. Il a aussi la caractéristique d'être traité en différé ; ce qui est problématique, puisque l'entreprise offrant le service n'a pas le même point de vue que l'utilisateur sur le temps nécessaire à une réponse. Pour l'utilisateur, 24 heures est un délai raisonnable entre la rédaction d'un courriel et la réception d'une réponse satisfaisante. Ce délai, qui devrait être la norme, est plutôt l'exception.

La réalisation d'un système de traitement automatisé des courriels constitue un défi technologique important, nécessitant l'apport de plusieurs domaines de l'intelligence artificielle. Dans cette thèse, nous présentons l'architecture d'un système de réponse automatisée basée sur le modèle de la question-réponse. Cette architecture est l'aboutissement d'un projet que nous avons abordé selon deux points de vue complémentaires : l'ingénierie et la science. Le projet est d'automatiser la réponse aux requêtes soumises par courriel au service de relation avec les investisseurs de la compagnie Bell Canada Enterprise (BCE). Nous avons conçu l'architecture à partir des connaissances existantes, et, simultanément, nous avons observé des phénomènes liés au problème à partir des expériences que nous avons réalisées pour construire le système.

1.1 Service de relation avec la clientèle

Les services de relation avec la clientèle (CRM - *Customer Relationship Management*) sont des composantes importantes pour les entreprises, la qualité des services offerts a un impact direct sur l'image d'une entreprise. La diversification des modes de communication pose un défi de taille parce que l'information est produite en plus grande quantité, elle circule plus rapidement et les usagers y ont accès plus facilement.

La combinaison de ces facteurs fait en sorte que les utilisateurs ont des attentes élevées envers ces services. Ils veulent une réponse de qualité à leur requête, et ce, dans un délai de temps très court. Le problème pour les entreprises devient alors difficile à gérer dans ce contexte. L'augmentation du nombre de communications et de la quantité d'informations a pour conséquence de surcharger les services d'information, de sorte qu'il est de plus en plus difficile de satisfaire convenablement les utilisateurs.

Il y a quelques années, les méthodes de communication privilégiées étaient le courrier postal et le téléphone. Dans le cas du courrier postal, son traitement n'était pas problématique car le client n'attendait pas une réponse instantanée. Les délais de transmission du courrier postal étaient connus et les communications devaient être prévues en conséquence. Par l'entremise d'une communication téléphonique, l'utilisateur espère être répondu instantanément. Le téléphone permet d'avoir une conversation directe avec un agent, mais il nécessite d'importantes ressources et il est réalisé en continu avec l'utilisateur. Une solution à ce problème consiste à utiliser un système de routage d'appels pour acheminer les appels des utilisateurs directement aux bonnes ressources. Les contraintes inhérentes à ces deux moyens de communication limitent l'achalandage à ces services en diminuant les attentes de la clientèle envers ceux-ci.

Dans le problème qui nous concerne, nous traitons les communications acheminées aux services de relation avec les investisseurs par le biais du courriel. Ces communications se font selon les mêmes principes que le service à la clientèle, mais l'aspect informatif de la tâche doit être considéré avec plus d'attention. Le service d'assistance (*help desk*) est une tâche qui donne une plus grande importance à l'information que le service à la clientèle. Le service d'assistance dans les entreprises a pour tâche de trouver des solutions aux problèmes provenant principalement de l'intérieur de l'entreprise. Puisqu'il constitue un centre névralgique d'information pour l'entreprise, ce service traite aussi des requêtes externes. Le service d'assistance doit être efficace et précis, il traite un ensemble de requêtes diversifiées autant par le contenu que par l'expertise nécessaire pour y répondre.

Les tâches réalisées par un service d'assistance sont diversifiées. Elles ne peuvent pas toutes être automatisées parce que certaines nécessitent impérativement l'intervention physique d'un expert. Dans cette thèse, nous proposons d'automatiser un sous-ensemble des tâches d'un service d'assistance, soit celui des processus informatifs

énoncés par une question. Les processus informatifs sont présents dans les services de références offerts dans les bibliothèques pour aider les utilisateurs dans leur quête d'information. Le service de référence est aussi offert en version électronique, dans ce cas nous pouvons le considérer comme un service de question-réponse utilisant la technologie de l'internet pour mettre l'utilisateur en contact avec un expert du domaine. L'expert n'a pas seulement la tâche de répondre à la question, il doit s'assurer de combler le besoin d'information de l'utilisateur en lui fournissant des références vers d'autres sources d'information (imprimée, électronique ou audiovisuelle). Le problème que l'expert tente de résoudre est différent de celui rencontré dans les systèmes experts traditionnels. Il n'est pas question ici de résoudre un problème comme une personne le ferait, ce qui est la définition du résultat d'un système expert, mais bien de fournir de l'information à un utilisateur à la manière d'un préposé du service des renseignements bibliographiques. Dans ce contexte, la difficulté du problème consiste à identifier le besoin d'information de l'usager, et l'utilisation qu'il fera de l'information. Le préposé devra parfois s'entretenir avec l'usager pour clarifier sa requête, ceci demandera une infrastructure informatique appropriée et des préposés formés pour réaliser cette tâche.

L'automatisation des services d'information est une façon d'optimiser les ressources pour endiguer le flot croissant de demandes. La mise en place de systèmes pour automatiser le traitement désengorgera les ressources, en traitant plus rapidement les requêtes répétitives et les questions simples, pour laisser les préposés répondre aux requêtes plus complexes nécessitant absolument une expertise pour le traitement. De plus, l'automatisation du service le rend accessible à toute heure de la journée sans contrainte géographique.

Par le traitement automatisé des requêtes, nous pouvons limiter les effets négatifs de l'augmentation du nombre de requêtes, tout en améliorant la qualité générale des réponses. Dans ces services, les préposées doivent continuellement actualiser leurs connaissances à cause de la nature dynamique de l'information. Et la disponibilité de l'information à partir de l'internet amène les utilisateurs à vouloir obtenir des réponses sans délai. L'automatisation du service permet de contourner ces problèmes par la synchronisation de l'information avec les connaissances du système. Par contre, il faut faire attention à la qualité des réponses fournies parce qu'un mauvais renseignement peut avoir de graves conséquences, telle l'aliénation du client envers l'entreprise.

Puisque ce type de service est habituellement réalisé par des personnes, l'automatisation doit s'intégrer au flot de traitement normal des requêtes. Le traitement d'une requête débute par une évaluation de la requête, qu'on achemine ensuite à la ressource appropriée. Cette dernière a la tâche de chercher l'information pour répondre à la requête. Mais la personne responsable de la requête doit parfois engager une conversation avec l'utilisateur dans le but de désambiguïser la requête ou de récupérer de l'information complémentaire nécessaire à l'identification de la réponse à retourner. L'étape finale du traitement de la requête consiste à rédiger une réponse à partir des sources documentaires dont le préposé dispose.

L'importance de l'automatisation du service de relation avec les usagers et la place qu'elle doit prendre dans les entreprises deviennent de plus en plus évidentes. Les entreprises constatent que les utilisateurs veulent plus d'informations, mais pas n'importe comment. L'utilisateur est moins sensible aux approches touchant la masse ; il veut seulement l'information qu'il désire et de la manière qu'il le désire. Pour satisfaire la clientèle, les communications avec les utilisateurs doivent être personnalisées.

Une solution entièrement automatisée serait envisageable si la précision du procédé d'automatisation s'approchait de la perfection. Puisqu'une telle précision n'est pas encore réalisable avec les méthodes actuelles de traitement de la langue et des connaissances, nous ne pouvons envisager d'automatiser complètement les services de relation avec la clientèle. Par contre, en exploitant des systèmes logiciels pour accomplir automatiquement certaines tâches, comme nous le présentons dans ce document, nous pouvons améliorer la qualité des services à la clientèle dans le cadre des services déjà offerts par l'entreprise. Cette solution permettra de diminuer le temps requis pour traiter une requête, d'où un service plus rapide et le traitement quotidien d'un plus grand volume de requêtes. De plus, les préposés les plus qualifiés pourront offrir un service personnalisé de meilleure qualité car les tâches les plus simples pourront être réalisées de façon assistée.

1.2 Traitement automatisé des courriels

Le courriel est l'application phare d'internet, il constitue l'activité principale des utilisateurs en ligne, c'est pourquoi nous considérons la réponse automatisée aux courriels pour automatiser les services d'information. Dans cette thèse, nous proposons

l'architecture d'un système pour répondre automatiquement aux courriels dans le but d'en améliorer le traitement. Cette architecture pourra aussi être utilisée pour traiter les communications électroniques en général. Nous avons limité notre cadre de recherche à l'automatisation du traitement des courriels spécifiques aux services d'information.

Le problème que nous avons considéré est celui d'automatiser la réponse aux courriels envoyés au service de relation avec les investisseurs de l'entreprise BCE, étudié dans le cadre du projet Merkure effectué au laboratoire de recherche appliquée en linguistique informatique (RALI). Deux approches, en plus de celle que nous présentons, ont été utilisées dans le projet Merkure pour solutionner ce problème. Julien Dubois, dans ses travaux de maîtrise [20], a évalué plusieurs approches pour classifier les courriels selon des caractéristiques communes. Luc Lamontagne, dans sa thèse de doctorat [40], a utilisé une approche basée sur le raisonnement à base de cas, pour mettre en correspondance une requête et un patron de réponse. Cette dernière approche calcule la similarité entre le message entrant et un ensemble de messages de la base de cas pour récupérer la réponse la plus appropriée.

Notre approche s'apparente aux travaux [38, 50] réalisé par Luc Plamondon, qui a développé le système de question-réponse factuel QUANTUM. Les résultats obtenus par ce système lors des conférences TREC sont comparables à la moyenne, nous pouvons ainsi le considérer comme un système de référence. Par contre, il ne traite que des questions factuelles énoncées directement. Ainsi, l'exploitation du système QUANTUM dans le cadre de la réponse automatisée aux courriels nécessite plusieurs changements pour rendre l'analyse de la question plus robuste et diversifier les types de questions traitées.

L'énorme quantité de courriels échangés entre les individus a créé un problème pour les grandes entreprises car elles doivent y répondre rapidement. Le traitement automatisé des courriels devient alors un problème qui doit être considéré. Les communications par courriels sont devenues si importantes et si abondantes qu'il est maintenant indispensable d'avoir des moyens pour les traiter automatiquement afin d'améliorer le service. Des études [5, 32] démontrent qu'il est essentiel pour les compagnies d'offrir un service de relation avec la clientèle qui traite les communications électroniques.

Différentes approches ont été proposées pour effectuer le traitement automatisé des courriels. Parmi les plus simples, il y a celles où l'expéditeur du courriel doit lui-même faire la classification de son courriel, pour l'envoyer au bon service, selon des adresses de courriels distinctes. Il y a également les systèmes d'auto-réponse, popularisés par les gestionnaires de listes de discussion (p. ex. LISTSERV[1], Mailman[2], Majordomo[3]), qui fonctionnent selon le principe d'activation par des mots-clés. Ces approches sont rudimentaires et elles ne peuvent pas être utilisées en pratique pour traiter efficacement le flot de communications par courriel d'un service avec la clientèle.

Une approche populaire dans les entreprises est l'utilisation de systèmes de gestion des courriels. Ces systèmes demandent une réponse manuelle mais contribuent à diminuer le temps de traitement des courriels en automatisant certaines fonctionnalités : la catégorisation des courriels, l'envoi d'accusé de réception, l'aiguillage des messages, la suggestion de réponses, l'intégration du système dans l'environnement de travail du préposé, l'archivage des courriels et la production de rapports statistiques et historiques. Les systèmes *Nomino Courriel*[4], *Kana Response*[5] et *eGain Mail*[6] sont des exemples de système de gestion des courriels pour automatiser une partie du processus de réponse.

Ces systèmes pour améliorer le traitement des courriels sont différents de ce que nous proposons. Dans notre cas, nous analysons le courriel pour chercher une réponse à partir de plusieurs sources. Nous ne voulons pas seulement retourner une réponse prérédigée à partir de la présence de mots clés ou traduire le message en une requête pour retourner les documents pertinents. Par l'architecture question-réponse, nous voulons identifier le but de la requête, en analysant en détail le courriel, et identifier précisément la réponse à la question provenant du courriel. En proposant une architecture générique pour solutionner le problème, nous voulons minimiser l'impact de la personnalisation du système par l'étape d'ingénierie des connaissances des systèmes actuels.

[1] `http://www.lsoft.com/`
[2] `http://www.list.org/`
[3] `http://www.greatcircle.com/majordomo/`
[4] `http://www.nominotechnologies.com/`
[5] `http://www.kana.com/`
[6] `http://egain.com/`

Le traitement automatisé des courriels est un domaine de recherche peu développé en dehors des systèmes commerciaux de gestion des courriels. La conférence scientifique *Conference on Email and Anti-Spam*[7] (*CEAS*), qui existe depuis 2004, est la seule dont le sujet est exclusivement l'étude du courriel. Par contre, la plupart des travaux concernant le courriel s'attarde aux polluriels (filtrage, techniques de pollupostage) ou étudie des problèmes sociologiques d'utilisation (identité, loi, politique). Le traitement automatisé des courriels reste un problème négligé, parce qu'il est difficile à réaliser et qu'il est difficile de convaincre les entreprises de partager leurs données (les courriels et les bases de connaissances permettant d'y répondre).

Nous avons recensé une approche du problème de réponse aux courriels par la « question-réponse » pour générer des réponses aux courriels envoyés dans une liste de discussion [70]. Les courriels qui y sont traités ressemblent à une conversation, il y a des courriels de questions et des courriels contenant les réponses. L'étape principale de leur système est l'extraction de phrases significatives dans les courriels à l'aide d'un pointage tenant compte de la présence de noms significatifs, de patrons d'expressions typiques d'une question et d'une pondération liée au nombre d'occurrences de la phrase dans les courriels réponse. Une fois les phrases significatives extraites, un calcul de similarité est effectué avec la base de courriels réponse pour déterminer quel courriel retourner comme réponse. L'approche qu'ils prennent pour solutionner le problème est similaire au raisonnement à base de cas étudié dans notre projet par Luc Lamontagne [40].

Le problème d'automatisation du traitement des courriels que nous considérons consiste à diviser le traitement en plusieurs étapes. Cette solution permet d'associer chacune des étapes de traitement à un module de l'architecture des systèmes de question-réponse. Dans ce travail, nous considérons seulement les problèmes d'automatisation concernant le traitement de la langue naturelle et l'intelligence artificielle en général. Il y a plusieurs problèmes que nous n'avons pas abordés, tels la mise en oeuvre technique du système de gestion des courriels et la conception d'interface permettant de traiter efficacement les requêtes.

[7] `http://www.ceas.cc`

1.3 Question-réponse

Le problème de la question-réponse fascine les chercheurs depuis que Turing a proposé de considérer la question « Can machine think ? » [65]. La tâche en question-réponse est de répondre à des questions formulées en langue naturelle. Le processus de question-réponse est complexe. Il demande d'abord la compréhension d'un besoin d'information exprimé en langue naturelle par une question, pour ensuite récupérer des ressources pertinentes pour la rédaction et la justification d'une réponse.

La question-réponse est une évolution importante des systèmes de recherche d'information fonctionnant à base de mots-clés et dont le résultat est une liste de documents. En question-réponse, l'utilisateur spécifie son besoin d'information par une question exprimée en langue naturelle. Ceci permet de définir et d'identifier clairement l'information qu'il cherche à récupérer. Le but de cette approche est de retourner une information plus pertinente à l'utilisateur. Il récupérera l'information qu'il demande avec un minimum d'effort sans avoir à fouiller à travers une liste de documents, comme c'est le cas avec un engin de recherche traditionnel.

Le développement de systèmes de question-réponse performants est essentiel pour exploiter les données électroniques disponibles à partir d'internet ou contenues dans des bases de données. Des méthodes pour exploiter ces ensembles de données sont connues, mais elles ne résolvent qu'une partie du problème, car leur portée est limitée. La question-réponse est une façon d'utiliser ces méthodes pour créer des systèmes facilitant l'exploitation de l'information par les utilisateurs.

Les premiers systèmes de question-réponse consistaient à offrir une interface pour interroger une base de données. Les systèmes les plus anciens sont BASEBALL [24] qui répond à des questions factuelles concernant des matchs de baseball de la ligue américaine et LUNAR [71] qui répond à des questions concernant des échantillons de sol lunaire. Ces systèmes sont maintenant connus par l'appellation d'interface en langage naturel à une base de données. Les systèmes experts s'apparentent aux systèmes de question-réponse, leur but étant de prévoir la réaction d'un expert face à une situation particulière. Le système expert le plus connu est MYCIN [62], qui répond à des questions concernant les diagnostics dans le domaine médical. Ces systèmes opèrent tous dans des domaines restreints, ainsi, une fois les connaissances modélisées, leur principal défi est de traduire une question en une requête pour récupérer la réponse.

L'information n'est pas toujours structurée, c'est pourquoi d'autres systèmes de question-réponse sont apparus pour traiter l'information textuelle. Les systèmes de question-réponse textuels ne nécessitent pas de bases de connaissances avec un format spécifique, ils tentent de découvrir la réponse en analysant des documents comme les articles provenant des journaux et des fils de presse. Le système QUALM [41] est un des premiers à traiter directement le texte dans le but de répondre à des questions concernant la compréhension d'une histoire. Ce système est à la base des travaux récents en question-réponse.

Depuis 1999, la recherche dans le domaine de la question-réponse est stimulée par la piste *Question Answering* de la *Text Retrieval and Evaluation Conference (TREC)* qui a débuté lors de TREC-8. Les systèmes précédents étaient difficiles à évaluer et à comparer parce que chacun travaillait à l'intérieur d'un domaine spécifique et le nombre de systèmes était restreint. L'évaluation réalisée lors des conférences TREC consiste à comparer les systèmes à partir d'un ensemble de questions à domaine ouvert. Les systèmes doivent répondre à un ensemble de questions factuelles à partir d'un corpus de documents composé d'articles de journaux et de fils de presse fournis par le *National Institute of Standards and Technology* (NIST). Cette initiative du NIST suscite un intérêt de la part de plusieurs équipes de recherche et la question-réponse est maintenant présente dans plusieurs domaines de recherche dont la recherche d'information et le traitement de la langue naturelle.

Les conférences TREC ont établi un cadre général pour concevoir les systèmes de question-réponse que la plupart des systèmes utilisent maintenant. L'architecture des systèmes de question-réponse issue de TREC est articulée autour de quatre étapes essentielles : l'analyse de la question, la sélection des documents candidats, l'analyse des documents et l'extraction de la réponse. Puisque cette architecture a démontré son efficacité, la plupart des systèmes s'en inspirent, mais les techniques pour chacune des étapes du traitement sont différentes d'un système à l'autre. Peu importe la manière dont le traitement est réalisé, le but des systèmes de question-réponse demeure toujours d'identifier les unités textuelles de la collection de documents correspondant au type d'information recherché.

L'architecture des systèmes de question-réponse est le cadre théorique principal de notre recherche. Les traitements que nous proposons pour solutionner le problème d'automatisation du service d'information s'intègre bien au cadre de la question-

réponse. L'analyse de la question est réalisée chaque fois qu'une demande arrive à un préposé pour déterminer l'action à prendre pour répondre au courriel. La personne détermine ensuite les sources qu'elle va utiliser pour satisfaire la requête et analyse les données qu'elle a récupérées. Selon l'action à prendre et le type d'information, il ne reste alors qu'à extraire la réponse et rédiger un message contenant l'information demandée par l'utilisateur.

1.4 Plan de la thèse

Nous débutons la thèse en présentant le problème de la question-réponse au chapitre 2. Nous y décrivons les problèmes abordés par la question-réponse, de même que son évolution depuis la mise en oeuvre des premiers systèmes. Par la suite, nous exposons l'architecture générique des systèmes de question-réponse factuelle. Dans le chapitre 3, nous mettons en contexte le problème de la question-réponse pour les services d'information et de gestion des relations avec les usagers. Ceci nous amène à présenter une architecture question-réponse pour solutionner le problème de la réponse automatisée aux courriels dans ce contexte.

Dans les chapitres 4 à 8, nous présentons les étapes pour répondre automatiquement aux courriels. Pour chaque étape de traitement, nous présentons un point de vue approprié pour traiter le problème nous concernant. Suite à la présentation de l'architecture et des composantes de traitement nécessaires à sa réalisation, nous revenons sur le contexte du problème et sur les manières d'exploiter les résultats que nous présentons dans cette thèse.

1.5 Contributions

Dans cette thèse, nous démontrons que nous pouvons réaliser un système de réponse automatisée aux courriels dans le cadre de l'automatisation des services à la clientèle à partir d'une architecture basée sur celle des systèmes de question-réponse. Une contribution de cette thèse réside dans l'originalité de l'approche, aucune méthode de traitement des courriels recensée dans la littérature n'aborde le problème sous l'angle de la question-réponse.

Nous avons établi que nous pouvons traiter les courriels envoyés à un service de relations avec la clientèle en suivant le même flot de traitement que celui couramment employé dans les systèmes de question-réponse factuelle. La différence entre la question-réponse et la réponse au courriel se situant dans la façon de traiter les données en fonction des particularités des domaines et des problèmes traités.

Nous avons étudié le corpus de courriels BCE-4 et avons extrait les caractéristiques les plus importantes à considérer pour le traitement automatisé des courriels dans le cadre du service de relation avec les investisseurs de la compagnie BCE. Suite à cette étude, nous avons développé une méthode à base de règles pour identifier les phrases qui sont des questions ou qui nécessitent une réponse dans un courriel.

Nous avons étudié différentes taxonomies de question liées à la problématique du triage dans les services d'information. Nous avons ensuite mis en relation ces taxonomies avec des niveaux de formalisation de la question. L'exploitation des taxonomies de questions nous permet de mettre en évidence le type de traitement linguistique que nous avons à réaliser pour analyser la question en fonction de l'information extraite de la taxonomie.

Nous avons réalisé un module de repérage des rôles sémantiques pour nous permettre d'identifier les rôles manquants d'une question et les relations exprimées par un verbe entre les entités. Ce module est réalisé en modélisant la tâche comme un problème de classification et en utilisant la méthode d'apprentissage supervisée SVM. Le repérage des rôles sémantiques peut ensuite être utilisé pour extraire des connaissances de documents textuels en fonction de l'ontologie du domaine que nous avons développée.

Nous avons aussi contribué aux développement du langage d'annotation du temps et des événements TimeML par le développement d'une représentation visuelle de l'annotation. De plus, nous avons intégré cette représentation graphique dans l'environnement d'annotation TANGO (TimeML Annotation Graphical Organizer).

Chapitre 2

Question-réponse

L'étude de la question-réponse et le développement des systèmes question-réponse sont réalisés pour faciliter l'exploitation et l'accès à l'information. L'information est omniprésente dans les entreprises de service, c'est une richesse qu'elles doivent exploiter et la question-réponse est une méthode pour le faire. La problématique abordée en question-réponse est de traiter une question formulée en langue naturelle pour en obtenir une réponse. La question est un mode de communication présent dans toutes les langues et une façon efficace d'exprimer un besoin d'information. Elle représente plus fidèlement l'information recherchée que les requêtes booléennes utilisées dans les systèmes de recherche d'information traditionnels.

La question-réponse est une approche générale qui nécessite la résolution d'un ensemble de problèmes présents dans plusieurs domaines de recherche. D'un point de vue philosophique et psychologique, la question et l'acte de réponse sont étudiés depuis plusieurs années. Les travaux de recherche réalisés pour automatiser les processus de question-réponse se retrouvent dans plusieurs branches de recherches : l'intelligence artificielle, le traitement de la langue naturelle, la théorie des bases de données, la recherche d'information et l'interaction homme-machine. Les systèmes pour traiter le problème de la question-réponse issus de ces domaines de recherche peuvent être classifiés en trois familles : les systèmes de base de données (pour interroger des données structurées), les systèmes de recherche de documents et les systèmes de compréhension d'histoire.

2.1 Évolution de la question-réponse

Les premiers systèmes utilisant l'approche de la question-réponse sont les systèmes de base de données, particulièrement ceux où la requête pour interroger la base de données est en langue naturelle. Ces systèmes retournent des réponses courtes provenant de la base de données interrogée, à partir d'une question où la sémantique est restreinte et précise. Les deux systèmes les plus connus opérant sur ce principe sont BASEBALL et LUNAR. Ces systèmes sont aussi connus sous l'appellation d'*interface en langue naturelle à une base de données*. Ils ne s'occupent pas explicitement du problème de l'extraction de la réponse, mais de la transformation de la question en une requête et de l'exécution de celle-ci.

Le système BASEBALL [24] répond aux questions concernant le score, les équipes, les endroits et les dates de parties de baseball. Les requêtes qu'il est capable de traiter sont simples, elles ne contiennent pas de conjonctions (*and, or, because, etc.*), ni de superlatif (*least, most, highest, longest, etc.*). La base de données interrogée est structurée autour d'un ensemble de couples attribut-valeur organisés dans une structure arborescente. La question est analysée automatiquement en transformant la phrase en une représentation attribut-valeur. Cette transformation est réalisée par une analyse partielle, en associant certains noms à des attributs à partir d'un dictionnaire, p. ex. *White Sox* est associé au couple `team:White Sox`. Les mots interrogatifs (*wh-words*), *who* et *where*, créent respectivement des entrées `team:?` et `place:?` dans la représentation de la question. Un mécanisme d'appariement de patron extrait ensuite la réponse en comparant la structure de la question à l'information dans la base de données. Lorsque les questions touchent plus d'un match de baseball, une procédure de recherche et de comptage est effectuée pour trouver la réponse. Ce système est limité par la quantité et la diversité de l'information présente dans la base de données. De plus, ses capacités de raisonnement restreignent le nombre de formes de question qu'il traite.

LUNAR [71] est un système qui permet aux géologistes d'interroger une base de données pour comparer et évaluer la composition chimique de roches lunaires recueillies lors de la mission Apollo-11. La base de données utilisée par LUNAR contient deux tables : une de 13 000 entrées contenant les analyses géologiques et une table de mots-clés servant d'index aux analyses. L'analyse des questions est réalisée par un

réseau de transition qui traduit la question dans le langage de requête de la base de données. La requête est construite en consultant un dictionnaire contenant de l'information syntaxique et morphologique et un ensemble de requêtes à la base de données, prérédigées et partiellement formulées. Les résultats obtenus par le système LUNAR ont montré que la question-réponse est une solution intéressante pour les utilisateurs voulant interroger une base de données. Les limitations du système proviennent du fait qu'il opère dans un domaine restreint et que ses traitements de données se font avec le dictionnaire et la base de données, qui représentent la majeure partie du travail de conception du système. Adapter ce système à un autre domaine nécessite presque autant de travail que la réalisation initiale.

L'acquisition des données n'est pas considérée comme un problème pour les interfaces en langue naturelle à une base de données, mais c'en est un en question-réponse puisque les données apparaissent sous la forme de textes non-structurés. Le système QUALM (*Question Answering Language Mechanism*) [41] est une étape importante dans le développement des systèmes de question-réponse. Il considère le problème de la question-réponse dans le contexte de la compréhension d'histoire. Le système utilise une représentation en graphes de dépendances conceptuelles (*Schank script*), de l'analyse de la question à l'élaboration de la réponse. La chaîne de traitement et la représentation des connaissances constituent un modèle psychologique du processus de question-réponse.

Le processus de question-réponse dans QUALM se divise en quatre étapes :

1. Catégorisation conceptuelle de la question ;
2. Inférence sur la question ;
3. Spécification du contenu ;
4. Heuristique pour récupérer la réponse.

La catégorisation des questions est réalisée sur une représentation de la question en graphes de dépendances conceptuelles, générée automatiquement à partir d'une grammaire. Dans cette catégorisation des questions, il est possible qu'une question appartienne à plusieurs catégories. Les catégories de questions considérées par QUALM se chevauchent parce que les catégories peuvent être définies par plusieurs propriétés. L'inférence sur la question est le procédé par lequel le contexte est pris en charge, ceci permet de restreindre la recherche de la réponse à une partie limitée du monde et reconnue par le système. La spécification du contenu détermine de quelle façon

la question sera répondue, le type de réponse et le niveau d'explication nécessaire pour que la réponse soit concluante. La récupération de la réponse est réalisée par une recherche à base d'heuristiques sur les histoires. Les types de recherches que QUALM réalise concernent les relations de cause à effet, la structure des scripts et la planification des actions. La recherche de la réponse est réalisée à partir des graphes conceptuels des histoires, qui capturent l'essentiel de l'information pouvant être interrogée, et d'une représentation du monde permettant d'associer une sémantique à la représentation. Le système QUALM identifie des problèmes à solutionner pour créer des systèmes de question-réponse à partir de données textuelles.

L'accès à une grande quantité de données a transformé le problème de la question-réponse. Retrouver l'information dans ces grands ensembles de données hétérogènes amène des problèmes nouveaux, mais possible à condition d'avoir : un système de recherche d'information performant, une requête représentant fidèlement le besoin d'information et le temps pour fouiller à travers une liste de documents considérés pertinents. Un problème auquel les systèmes de recherche d'information sont soumis provient du fait que la requête formulée par l'utilisateur doit correspondre à l'information qu'il recherche. L'utilisateur n'est pas toujours en mesure d'identifier ce qu'il recherche par sa requête car s'il le savait, il consulterait immédiatement les bonnes ressources. Le système START (*SynTactic Analysis using Reversible Transformations*) du MIT [34, 33] est le premier à proposer aux utilisateurs d'internet d'entrer une requête sous la forme d'une question et d'obtenir une réponse courte. Il permet d'obtenir rapidement une information qui nécessiterait habituellement l'aide d'un expert. La couverture du système est limitée aux questions factuelles concernant la géographie, le cinéma, les personnalités, les définitions de dictionnaires et généralement, les connaissances encyclopédiques. AskJeeves[1] est la première manifestation publicisée d'un système question-réponse commercial. Ce système d'information accepte des requêtes sous la forme d'une question pour ensuite retourner des réponses courtes ou une liste de ressources pertinentes lorsqu'il ne connaît pas la réponse, à la manière d'un engin de recherche, que l'utilisateur se servira pour trouver la réponse.

Le procédé de question-réponse est considéré selon trois aspects généraux de traitement, cette division permet de concevoir individuellement des approches pour construire des systèmes complets plus performants. Les aspects du traitement dans

[1] `http://www.ask.com/` (Jeeves est maintenant à la retraite)

les systèmes de question-réponse contemporains sont :

1. l'identification du besoin d'information d'un utilisateur par une interaction homme-machine ;
2. la sélection des sources d'information en fonction du besoin de l'utilisateur ;
3. l'évaluation des sources d'information en fonction du besoin d'identifier précisément l'information demandée et d'expliquer pourquoi elle correspond à ce qu'il recherche.

Les méthodes utilisées en question-réponse sont à l'intersection de plusieurs domaines de recherche : le traitement de la langue naturelle, la recherche d'information et l'interaction entre l'homme et la machine.

2.2 Question-réponse à TREC

La question-réponse factuelle est la piste de recherche la plus étudiée du domaine de la question-réponse. La popularité de cette piste de recherche est alimentée par les besoins des utilisateurs qui veulent avoir des réponses à leurs questions sans avoir à évaluer la pertinence des pages web retournées par les engins de recherche. La question-réponse factuelle traite les questions dont les réponses sont exprimées comme des faits simples.

L'intérêt des chercheurs en extraction d'information et en recherche d'information pour la question-réponse provient de la conférence TREC (*Text Retrieval and Evaluation Conference*) [69], organisée par le NIST, qui agit comme catalyseur des communautés de recherche dont le but commun est de solutionner les problèmes de la question-réponse. Cette conférence, où les systèmes de recherche d'information sont évalués, permet aux chercheurs de comparer l'efficacité des approches utilisées pour différentes tâches : recherche de documents, filtrage, recherche de documents vocaux, recherche de documents vidéo et question-réponse. La tâche de question-réponse consiste à traiter une question factuelle et à retourner une réponse sans intervention humaine au cours du processus. Les problèmes abordés par les systèmes de question-réponse lors des conférences TREC orientent les pistes de recherches que les groupes participants explorent.

Le projet Merkure, dans lequel se situe les travaux que nous présentons, a été influencé dès ses premières heures par la conférence TREC-8 (1999), première année où la question-réponse fut considérée. Le choix de l'approche question-réponse pour la réponse automatisée aux courriels a été fait suite aux résultats encourageants publiés lors de TREC-8 et TREC-9 et à la similarité entre la tâche rapportée et celle à réaliser dans notre projet. En présentant l'évolution de la tâche question-réponse lors des conférences TREC nous pouvons mettre en perspective les difficultés rencontrées dans l'élaboration d'un système de réponse automatisée aux courriels où nous devons fournir des réponses exactes à des questions. Nous débutons par la présententation la conférence TREC-8 et nous irons jusqu'à TREC-2004, dernière conférence ayant influencé nos travaux.

TREC-8

La tâche question-réponse évaluée à TREC a évolué depuis TREC-8 en 1999 [67]. Lors de la première conférence, la collection de documents était composée de 528 000 articles de journaux et de fils de presse à partir desquels les réponses à 200 questions devaient être extraites. Les questions factuelles ont été créées manuellement pour l'évaluation à partir de la collection de documents, p. ex. *How many calories are there in a Big Mac ?*, toutes les questions avaient donc une réponse dans la base de documents. Les réponses aux questions étaient nécessairement courtes puisque les participants devaient fournir deux listes de cinq réponses d'au plus 50 et 250 caractères pour chacune d'elles. Les questions s'inspiraient de questions des archives de requêtes de FAQFinder [10].

L'évaluation des systèmes à TREC est faite en différé, les participants reçoivent une liste de questions et ils ont une semaine pour retourner les résultats obtenus. L'évaluation des réponses est réalisée manuellement par des examinateurs du NIST. La mesure d'évaluation des systèmes est le *mean reciprocal rank* (MRR). Le MRR d'une question est l'inverse du rang auquel on retrouve la bonne réponse dans la liste de réponses que le système donne en sortie. Lorsque la bonne réponse est en première position le MRR est $1/1 = 100\%$, lorsque la réponse est en $4^{\text{ème}}$ position, le MRR est seulement $1/4 = 25\%$. Le MRR pour un système est la moyenne du MRR calculé sur chacune des questions de l'ensemble de questions évaluées. Cette méthode d'évaluation donne un avantage aux systèmes donnant de bonnes réponses au début de la liste, dans les premières positions.

Lors de TREC-8 une réponse était considérée bonne lorsque la chaîne de caractères contenait la bonne réponse. Et le numéro de document, agissant comme justification de la réponse, n'était pas considéré pour évaluer la réponse. Vingt équipes ont participé à cette première évaluation, la meilleure équipe (Cymfony Inc.) obtenant un MRR de 0.660 avec 54 questions ne contenant pas la bonne réponse lors de l'évaluation sur 50 caractères. L'équipe qui est arrivée en seconde position (Southern Methodist University) a obtenue un MRR de .646 avec 44 questions ne contenant pas la bonne réponse lors de l'évaluation sur 250 caractères.

TREC-9

La tâche de question-réponse évaluée à TREC-9 (2000) est essentiellement la même que celle de TREC-8, les réponses sont toujours présentes dans le corpus et ont une longueur maximale de 50 caractères. La collection de documents est plus volumineuse (979 000 documents) et le nombre de questions est passé de 200 à 693. De ces 693 questions, 500 sont des questions originales, les 193 autres sont des réécritures de 54 différentes questions parmi les 500 originales. Les questions proviennent de questions soumises à l'encyclopédie en ligne Encarta de Microsoft et à l'engin de recherche Excite. L'évaluation des systèmes a été modifiée en introduisant le concept de réponse justifiée, par lequel une réponse est bonne seulement si le document associé à la réponse permet de justifier la réponse. La mesure MRR a été conservée pour comparer les systèmes. Les scores obtenus par les systèmes sont plus bas que ceux de l'année précédente, mais dans l'ensemble, les systèmes donnent de meilleurs résultats parce que : les questions sont plus difficiles, le nombre de questions à répondre est plus grand et la justification des réponses par un numéro de document élimine des réponses qui étaient bonnes auparavant. Le système Falcon de Southern Methodist University a obtenu les meilleurs résultats (MRR de 0.58 et 229 (34%) questions sans réponses), loin devant les autres, autant avec les réponses de 50 caractères que de 250.

TREC-2001

Les nouveautés apportées à la tâche de question-réponse lors de TREC-2001 sont : de limiter les réponses à 50 caractères, d'inclure des questions dont la réponse n'est

pas dans la collection de documents et d'inclure des questions où la réponse est éparpillée dans plusieurs documents. Pour les questions n'ayant pas de réponse, les systèmes doivent retourner NIL comme réponse. L'évaluation principale consiste à répondre à 500 questions dont 135 sont des définitions. Deux nouvelles évaluations y sont effectuées, la première concerne 25 questions devant retourner une liste comme réponse, la seconde est de répondre à des questions en suivant un contexte de questions successives, à la manière d'une conversation.

TREC-2002

Lors de TREC 2002, on ne recherche que les réponses exactes et les listes de réponses. La tâche de question-réponse contextuelle est abandonnée puisque la compétence des systèmes à répondre aux questions contextuelles ne dépendait pas de leur habilité à traiter le dialogue sous-entendu, mais des capacités des systèmes à répondre aux types de questions. La réponse exacte ne doit pas contenir d'information ne contribuant pas à la réponse comme ce pouvait être le cas auparavant avec une chaîne de 50 caractères. Une nouvelle mesure de comparaison est utilisée, le score pondéré de confiance, qui permet de tester l'aptitude du système à détecter les bonnes réponses. Les réponses retournées par les systèmes doivent être modifiées pour accommoder cette nouvelle mesure. Les systèmes ne retournent plus cinq réponses à une question mais une seule et les réponses aux 500 questions doivent être ordonnées en ordre décroissant de certitude. Le score pondéré de confiance se définit en fonction de l'ensemble de questions $\mathcal{Q}$ et du nombre de bonnes réponses aux i premières questions,

$$\frac{1}{\|\mathcal{Q}\|}\sum_{i=1}^{\|\mathcal{Q}\|}\frac{\|\textit{bonnes réponse parmi les } i \textit{ premières}\|}{i}$$

Cette mesure a pour effet de récompenser les systèmes qui trouvent les bonnes réponses et qui sont capables d'identifier les réponses dont ils ont le plus confiance qu'elles soient bonnes. L'augmentation de la difficulté de la tâche question-réponse à TREC 2002 a eu pour effet de diversifier les approches adoptées par les systèmes pour résoudre les problèmes que chacun a vu apparaître avec la nouvelle façon d'évaluer les systèmes.

TREC-2003 et TREC-2004

Les évaluations de la question-réponse réalisées lors de TREC 2003 et TREC 2004 [68] sont similaires, les réponses aux questions que les systèmes doivent maintenant fournir sont un mélange de réponses factuelles courtes, de listes et de définitions. Lors de TREC 2004, les questions ont été groupées en séries et ordonnées de sorte que certaines questions doivent avoir la réponse des autres questions de la série pour être répondues. La difficulté de la tâche est plus grande qu'auparavant, les systèmes qui performaient bien continuent de bien faire mais les systèmes utilisant de nouvelles approches ont tendance à se retrouver en queue de peloton. Les nouvelles approches ayant tendance à solutionner des problèmes spécifiques dont la solution n'est pas mise en valeur par le score de confidence pondéré.

2.3 Architecture des systèmes de question-réponse

L'architecture des systèmes de question-réponse traitant les questions factuelles est similaire dans la plupart des implémentations. Cette situation est en partie redevable aux évaluations menées lors des conférences TREC qui a eu pour effet de propager les méthodes qui fonctionnent le mieux à travers le domaine de recherche. L'architecture générale d'un système de question-réponse factuel utilisant un corpus textuel comporte cinq étapes de traitement (fig. 2.1). Nous avons ajouté une sixième étape à cette architecture, la génération de la réponse, qui est nécessaire lorsque la question-réponse est utilisée pour répondre à un utilisateur.

1. **Le prétraitement des documents de référence** permet de transformer les données textuelles en connaissances utilisables pour répondre aux questions.
2. **L'analyse de la question** a pour fonction de comprendre le sens de la question et de déterminer le type d'information recherchée.
3. **La sélection des documents candidats** extrait un sous-ensemble de documents contenant possiblement une réponse à la question ou de l'information pertinente devant être analysée pour en déduire une réponse.
4. **L'analyse des documents candidats** est effectuée pour extraire de l'information provenant des documents candidats pour préciser la question ou ajouter des faits qui aideront à trouver la réponse.

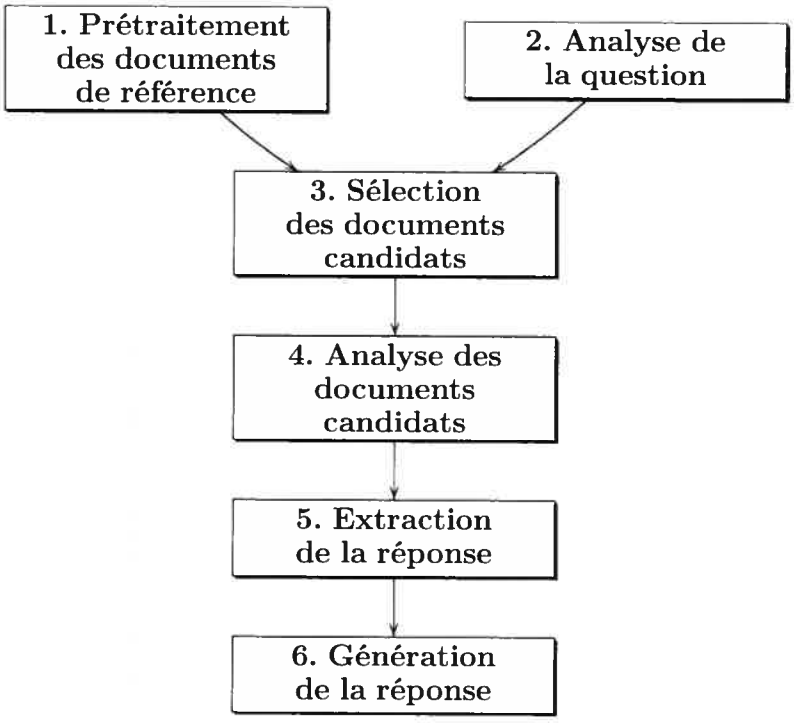

FIG. 2.1 – Architecture générale d'un système de question-réponse

5. **L'extraction de la réponse** est réalisée à partir d'une représentation de la question et d'un mécanisme d'appariement, permettant d'extraire un ensemble de réponses candidates qui sont ensuite évaluées pour en déterminer la réponse la plus vraisemblable.
6. **La génération de la réponse** est réalisée en fonction du contexte de l'utilisateur, à partir des réponses extraites du corpus. Cette étape n'est pas présente dans les systèmes évalués à TREC mais elle est essentielle pour l'utilisation courante des systèmes de question-réponse.

Les composantes d'un système de question-réponse s'organisent autour de trois approches du problème : recherche d'information, extraction d'information ou traitement de la langue naturelle. Chaque système de question-réponse est influencé par le champ de compétence du groupe de recherche responsable de sa conception. Lorsque le problème de la question-réponse est abordé sous l'angle de la recherche d'information, l'effort porte sur la recherche du document ou du passage le plus pertinent pouvant répondre à la requête. L'approche par extraction d'information met l'accent sur l'identification de l'information, principalement en utilisant des patrons à

trous pour extraire le contenu informatif des documents. Tandis que la considération du problème sous l'angle du traitement de la langue naturelle consiste à améliorer les techniques de recherche et d'extraction d'information en analysant les requêtes et les documents plus en profondeur. Les systèmes actuels doivent intégrer les trois approches pour répondre aux questions en offrant des performances acceptables.

2.3.1 Prétraitement des documents de référence

Le corpus de documents de référence doit être préparé pour pouvoir l'exploiter efficacement. Le prétraitement des documents de référence consiste principalement à indexer les documents par un moteur de recherche. C'est aussi l'occasion d'aller récupérer directement l'information dans les textes et d'en faire une base de connaissances.

Même si les documents candidats sont sélectionnés par un système de recherche d'information conventionnel, le prétraitement des documents accélérera la tâche d'analyse des documents candidats. L'information extraite ou ajoutée par le prétraitement est conservée en parallèle parce que l'ajout de cette information à la base d'indexation des documents biaiserait la sélection des documents candidats. L'utilisation d'une étape de prétraitement évoluée augmente le temps global de calcul, mais permet d'obtenir une représentation logique ou sémantique du texte beaucoup plus riche que si elle avait été réalisée simultanément à l'analyse des documents candidats. Voici donc quelques prétraitements implémentés dans les systèmes de question-réponse :

- l'analyse de surface extrait des patrons identifiant des éléments importants du texte ;
- l'étiquetage des parties du discours (*POS*) permet d'obtenir le rôle grammatical de chaque mot sans faire une évaluation syntaxique complète ;
- la reconnaissance d'entités nommées identifie les composantes intéressantes du texte comme les noms de personne, de ville, de compagnie, etc. ;
- le découpage (*chunking*) réalise des groupements de mots selon leur relation fonctionnelle dans la phrase ;
- la dérivation de représentations logiques sert à réaliser des inférences sur les représentations ;
- l'annotation des termes par des marqueurs sémantiques selon l'information que

l'on veut aller chercher ;
- l'extraction de relations sémantiques entre les entités.

Le système START [33] traite les documents du corpus en les ajoutant à Omnibase [34], une base de données virtuelle. L'approche d'Omnibase est d'aller récupérer l'information sur le web et de la structurer sur la base du modèle *objet–attribut–valeur*. L'identification de l'information fonctionne selon des patrons qui annotent l'information. L'annotation est réalisée par l'ajout de méta-données à la ressource en liant le triplet d'information au texte original. Par exemple, l'annotation des données telle que proposée par le W3C pour implémenter le web sémantique est un type de prétraitement des données où les auteurs des documents annotent eux-mêmes les informations.

2.3.2 Analyse de la question

L'analyse de la question est un élément primordial d'un système de question-réponse. Si cette analyse est erronée, les chances de trouver une bonne réponse à la question sont sérieusement compromises. L'analyse de la question doit déterminer le type d'information recherchée. Lorsque la question est ambiguë, l'analyse doit utiliser des stratégies de recherche pour mettre la question dans un contexte permettant de la désambiguïser.

La classification des questions restreint les possibilités d'analyses en utilisant des taxonomies de questions particulières à chaque système. Elle est réalisée soit par appariement de patrons ou par apprentissage. Certains systèmes utilisent des méthodes plus complexes, Hermjakob [28] réalise une analyse syntaxique complète de la question et applique ensuite plusieurs règles sur l'arbre de dérivation pour classifier la question.

Parallèlement à la classification, les mots de la question sont soumis à une analyse morpho-syntaxique leur associant une étiquette d'une partie du discours. Les étiquettes sont ensuite utilisées pour transformer la question en une requête pour la sélection des documents candidats. La requête est formée en utilisant l'information relative à la classification de la question et au but identifié, à partir de formes de requêtes prédéfinies.

Les systèmes plus sophistiqués transforment les questions dans des représentations sémantiques à la manière du système QUALM. Ces systèmes exploitent une grammaire pour générer une dérivation syntaxique utilisée pour représenter la question. Cette représentation de la question n'est pas utilisée lors de la recherche de documents candidats, mais lors de l'identification de la réponse.

Dans cette étape de traitement, aucune méthode ne semble être significativement plus performante. L'important est de constater que les méthodes d'analyse des questions sont liées aux traitements réalisés lors des étapes subséquentes. L'analyse de la question doit fournir une information et une représentation pertinente de celle-ci pour identifier la réponse. L'analyse de la question est une tâche difficile, elle constitue un goulot d'étranglement à l'élaboration de système plus performant, mais peu de systèmes ont considéré cette étape en priorité.

2.3.3 Sélection des documents candidats

La sélection des documents est habituellement réalisée à l'aide d'un système de recherche d'information. Le but de cette étape étant de restreindre le nombre de documents dans lesquels la réponse est cherchée et non pas de trouver la réponse. L'information recherchée en question-réponse est beaucoup plus spécifique qu'en recherche d'information traditionnelle. Pour identifier l'information précisément, la sélection des documents se fait avec des requêtes booléennes pouvant être ajustées au besoin d'information ou par l'extraction de passages de textes. L'approche d'extraction de passages est fondée sur la prémisse que l'information recherchée pour répondre à une question est présente dans certaines parties restreintes d'un document. Les passages retournés par l'extraction ont l'avantage d'être plus courts que les documents complets et sont donc plus simples à analyser en détail.

La sélection des documents par la recherche d'information est un domaine de recherche mature qui possède plusieurs méthodes ayant démontré leur efficacité à récupérer des documents pertinents à une requête. Les systèmes de recherche d'information se différencient par la façon dont les documents sont indexés, mais principalement par leur façon de calculer la similarité entre une requête et un document. Les modèles couramment utilisés sont les modèles booléen, vectoriel et probabiliste. La performance d'un engin de recherche est influencée par la méthode qu'il implémente

et surtout par la mise au point des paramètres qui doivent être ajustés selon la tâche à réaliser.

Les requêtes traitées par les systèmes de recherche d'information ne sont pas adaptées à la question-réponse parce que la recherche retourne les documents similaires à la requête qui est une question. L'utilisation de la question comme requête ne discrimine pas avec assez de précision les documents recherchés par la question, la réponse ne sera trouvée que lorsqu'elle sera formulée dans une forme similaire à la question. L'analyse de la question permet de construire une requête pour identifier les documents pertinents à l'extraction de la réponse. Dans le contexte de la recherche d'information l'analyse de la question s'apparente à la réécriture de la requête. L'extraction des documents pertinents est une étape aussi importante que les précédentes puisqu'elle influence elle aussi la performance des étapes suivantes.

2.3.4 Analyse des documents candidats

L'analyse des documents candidats consiste à fouiller l'information retournée par l'étape de sélection des documents pour identifier des phrases ou des morceaux de phrases correspondant au type d'information recherché. Les méthodes d'analyse des documents sont les mêmes que celles utilisées en extraction d'information, généralement c'est l'identification des entités nommées et des relations sémantiques. L'identification des entités nommées est utilisée pour identifier le type sémantique de l'information contenue dans les documents candidats. L'identification des relations sémantiques sert à interpréter les liens entre les entités.

La tâche minimale à exécuter lors de cette étape, pour qu'un système fonctionne bien, est l'identification des entités nommées, c'est-à-dire extraire et classifier les noms de lieux, de personnes, de compagnies, les adresses, les numéros de téléphone, les liens URL et les mesures. D'autres classes et sous-classes peuvent être ajoutées, celles-ci n'étant que les plus communes. Les autres traitements pouvant être exécutés sur les documents sont sensiblement identiques à ceux qui peuvent être appliqués lors du prétraitement de la collection, soit : le découpage en phrase, l'annotation des parties du discours (POS), le « chunking », etc. Lorsque la phrase contient une entité correspondant au type sémantique de l'information recherchée, la phrase est ajoutée à la liste des candidats à considérer pour l'identification de la réponse.

La phrase candidate est ensuite liée à la question pour déterminer si elle correspond à la structure de la question. La liaison de la phrase à la question se fait en analysant l'arbre de dérivation syntaxique ou les relations de dépendance dans la phrase. Les informations pertinentes de la phrase sont généralement les sujets de clauses relatives qui sont associées au groupe nominal de la question. Une autre façon de lier la phrase à la question est d'utiliser des patrons associés à un type d'information recherché. Lorsque des dates de naissance ou de mort sont considérées, cette approche est très utile car cette information se présente souvent sous la forme d'un patron, p. ex. `Abraham Lincoln (1809-1865)`, représenté par `NAME (YEAR_BIRTH-YEAR_DEATH)`. Il faut parfois utiliser des méthodes plus complexes nécessitant des ressources lexicales plus évoluées pour déterminer le sens des mots utilisés dans les questions. Ces méthodes sont essentielles pour répondre à certaines questions car plusieurs raisonnements faits inconsciemment par une personne deviennent impossibles à réaliser [27].

Lorsqu'il est impossible de lier la question à une phrase candidate, la stratégie est alors de prendre les phrases dont la mesure de similarité est la plus grande. La mesure de similarité pose la contrainte que les mots de la question doivent être présents dans la phrase ou dans celle précédant ou suivant celle-ci.

2.3.5 Extraction de la réponse

L'extraction de la réponse est réalisée à partir d'une représentation de la question et des segments de textes retournés par l'analyse des documents candidats. La réponse retournée est rarement unique, une liste de réponses ordonnées selon une mesure de confiance est généralement plus appropriée. La sélection de la réponse est réalisée de plusieurs manières différentes dans les systèmes de question-réponse, selon la représentation de la question, la sélection des candidats et les ressources utilisées par le système.

Une manière de sélectionner les réponses consiste à contraindre les candidats à répondre à des critères de sélection. Le premier critère utilisé est de comparer le type ou la catégorie sémantique de la réponse attendue. Les réponses candidates sont ensuite comparées entre elles pour identifier les réponses les plus fréquentes. La comparaison des fréquences est habituellement effectuées sur le sous-ensemble de documents candidats sélectionnés, mais certains systèmes utilisent le corpus complet pour calculer

la fréquence d'apparition de la réponse avec les termes de la question. Le web est aussi utilisé pour confirmer la sélection de la réponse en utilisant la fréquence d'apparition de la réponse. La comparaison des réponses selon leur fréquence fonctionne à condition d'admettre l'hypothèse qu'un énoncé plus fréquent est celui qui correspond le plus fidèlement à la réponse. Dans le problème que nous considérons cette hypothèse ne peut pas être exploitée parce que nous travaillons à partir d'une petite collection de documents spécifiques au domaine, contrairement à la question-réponse traditionnelle qui exploite de gros corpus de documents.

La méthode précédente pour identifier une réponse est appropriée pour les approches de la question-réponse traitant les données textuelles par des approches en surface. Lorsque le système utilise une représentation plus riche en connaissances (prédicats logiques, annotations sémantiques, relations <sujet,verbe,objet>), des techniques inspirées des interfaces en langue naturelle aux bases de données et des modèles d'inférence sont utilisées. Cette représentation permet d'utiliser des contraintes plus compliquées et d'effectuer certains raisonnements simples pour identifier les réponses. L'inconvénient de cette méthode est de faire diminuer le taux de rappel en faveur de la précision, le temps de calcul peut aussi être prohibitif pour certaines questions en fonction de la taille de la base de connaissances qui est utilisée.

La réponse peut aussi être identifiée à partir d'heuristiques construites manuellement correspondant à des types de réponses prédéterminées. Ces heuristiques performent bien sur les types de questions pour lesquelles elles ont été construites, c'est une façon efficace de sélectionner la réponse mais elle est vulnérable lorsque la question ne peut être associée à une heuristique.

À titre d'exemple le système de question-réponse QUANTUM [50, 51] conçu au laboratoire RALI de l'Université de Montréal utilise ce type d'heuristique sous l'appellation de fonctions d'extraction. Les fonctions d'extraction sont définies dans une hiérarchie selon le type attendu de la réponse. Les réponses retournées par les fonctions sont ordonnées selon une mesure de confiance calculée par la fonction. Les onze fonctions d'extraction de QUANTUM sont divisées en cinq catégories :

1. **Les fonctions de hiérarchie** sont la *définition* et la *spécialisation*, elles sont basées sur les relations d'hyponymie et d'hyperonymie que nous retrouveons dans WordNet. Ces fonctions d'extraction utilisent l'extraction d'entités nommées et les relations d'hyponimie et d'hyperonymie pour trouver les réponses.

2. **Les fonctions de quantification** sont la *cardinalité* et la *mesure* qui retournent des nombres. La fonction *cardinalité* est applicable lorsque la question porte sur une quantité qui peut être comptée. La fonction *mesure* identifie les nombres qui expriment une quantité mesurable et mesurée.
3. **La fonction de caractérisation** est l'unique fonction *attribut* qui extrait des informations concernant les attributs de l'objet identifié par l'entité nommée de la question.
4. **Les fonctions de complétion de concept** sont : *personne*, *temps*, *lieu* et *objet*. Ces fonctions sont utilisées pour traiter les questions utilisant un pronom interrogatif pour faire référence à l'entité recherchée.
5. **Les autres fonctions** sont la *manière* et la *raison* qui n'appartiennent pas à une catégorie particulière.

Les fonctions d'extraction de QUANTUM sont, pour la plupart, des expressions régulières sur les mots et leur étiquette grammaticale. Certaines fonctions nécessitent des paramètres provenant de l'analyse de la question, ceci permet d'avoir de l'information supplémentaire pour trouver la réponse.

2.3.6 Génération de la réponse

Nous n'avons pas étudié la génération de la réponse en profondeur parce que nous n'avons pas abordé ce problème dans notre architecture. De plus, nos travaux s'inspirant de la question-réponse factuelle, les réponses générées par ces systèmes sont des réponses courtes ou des listes de réponses, elles ne sont jamais formatées pour être retournées à un utilisateur. La génération de la réponse est un problème exigeant et c'est le sujet d'une partie des travaux de Luc Lamontagne [40], dans notre projet concernant l'automatisation des services de relations avec les investisseurs de la compagnie BCE par la réponse automatisée aux courriels.

2.4 Conclusion

La question-réponse est l'approche que nous avons privilégiée pour solutionner le problème de la réponse automatisée aux courriels. Les problèmes que nous rencontrons pour automatiser la réponse aux courriels partagent des caractéristiques

similaires aux trois familles de systèmes de question-réponse que nous venons de présenter. Nous devons répondre à des questions nécessitant l'interrogation d'une base de données lorsqu'elles concernent des données structurées ; trouver des documents pertinents à la question lorsque l'information demandée est partiellement définie ou la réponse ne peut être formulée succinctement ; « comprendre » l'information textuelle pour retourner des réponses courtes à partir d'une ou plusieurs sources.

Les systèmes de question-réponse factuelle se ressemblent puisqu'ils utilisent une architecture similaire. Le système QUANTUM n'est pas différent des autres, en utilisant des ressources plus élaborées pour extraire l'information et identifier les relations, les performances du système seraient meilleures. Nous pouvons exploiter l'expérience acquise lors du développement du système de question-réponse factuelle QUANTUM pour la réponse automatisée aux courriels, en adaptant et en améliorant les étapes de traitement les plus pertinentes à la réponse automatisée aux courriels. Globalement, le système QUANTUM ne peut pas être utilisé directement pour les courriels parce qu'il n'est pas construit pour analyser les questions contenues dans les courriels et l'information recherchée dans les questions s'extrait difficilement avec les fonctions d'extraction définies dans le système. Dans le prochain chapitre, nous reprenons les étapes de traitement décrite brièvement pour situer le problème de la réponse automatisée aux courriels dans le contexte du problème de la question-réponse.

Chapitre 3

Architecture du système

Dans cette thèse, nous soutenons que l'architecture et les méthodes de la question-réponse peuvent être utilisées pour solutionner le problème de la réponse automatisée aux courriels dans le cadre des services d'information et de la gestion des services de relation avec la clientèle. Dans le chapitre précédent, nous avons présenté les pistes qui ont été abordées pour solutionner la question-réponse, ainsi que l'architecture générale des systèmes de question-réponse textuels répondant à des questions factuelles. Dans ce chapitre, nous identifions les caractéristiques du problème de question-réponse propre à notre problème. Le contexte du problème nous permet de définir une architecture pour la création d'un système de réponse automatisée aux courriels, similaire à celle d'un système de question-réponse.

3.1 Question-réponse pour les services d'information et de gestion des relations avec les usagers

La question-réponse, les services d'information et les services de relation avec les usagers ont plusieurs points en commun, mais aussi plusieurs différences faisant en sorte que la question-réponse doit être adaptée aux problèmes spécifiques reliés à ces services. Les fonctionnalités qu'un système de question-réponse doit comporter se décrivent en fonction des différences entre les problèmes de la question-réponse et ceux observés dans l'offre d'un service. L'architecture d'un système de question-

réponse dans un cadre applicatif est réalisée en considérant les besoins des utilisateurs, l'adaptabilité du système et l'intégration dans le traitement du service avec les usagers.

3.1.1 Sources d'informations

Les questions traitées par un service d'information sont généralement limitées à un domaine d'application où les termes et la manière de présenter les faits lui sont spécifiques. Une architecture question-réponse pour ce problème doit donc compter sur des ressources linguistiques adaptées et sur un ensemble de connaissances particulier au domaine d'application. Ces ressources, définissant un modèle du domaine, sont utilisées de la première étape : l'analyse de la question, à la dernière : la génération de la réponse.

Le domaine de traitement du service d'information est lié à celui de l'entreprise qui offre le service et aux connaissances visées par les requêtes considérées. Par exemple, dans notre projet, le service d'information traite les relations avec les investisseurs, il doit donc avoir une connaissance du milieu des investisseurs (c'est-à-dire comprendre les termes de la finance, la comptabilité, l'investissement) et une compréhension des données liées aux modèles du domaine, mais dans le contexte particulier de l'entreprise. Ceci fait en sorte qu'un système de question-réponse doit exploiter une multitude de sources d'informations.

La première source de références considérée pour les systèmes de question-réponse est le site web de l'entreprise contenant l'information rendue accessible au public par l'entreprise. Cette source d'informations est du même type que celle utilisée pour les systèmes de question-réponse factuels. Une autre catégorie d'information disponible pour la question-réponse est le web invisible, constitué de l'information contenue dans des bases de données accessibles par une interface de recherche. La récupération de cette information nécessite une description exacte de ce qu'on cherche en fonction de la structure de la base de données. Il y a aussi l'information qui n'est pas disponible publiquement, pour des raisons de politiques d'entreprise ou de confidentialité de l'information. Cette dernière source d'informations doit aussi être exploitée car les services d'information desservent à la fois des requêtes internes et externes.

La tâche de question-réponse qui doit être résolue pour répondre à ce problème est limitée par le domaine de traitement du service d'information, mais elle ne correspond pas exactement à la définition de la question-réponse dans un domaine restreint parce que les questions peuvent porter sur plusieurs aspects concernant plus d'un domaine précis. La division des domaines de connaissance est relative à l'organisation de l'information dans l'entreprise.

Les questions posées au service d'information nécessitent des raisonnements complexes et diversifiés, elles ne sont pas seulement factuelles. Dans le cas du service de relation avec les investisseurs, les requêtes touchent des dossiers d'usagers, des documents à récupérer, des précisions concernant des documents publiés par l'entreprise, des méthodes pour solutionner un problème, des confirmations d'informations, etc. Ces demandes dépassent largement le cadre de la question-réponse factuelle. Un système de question-réponse ne peut pas traiter tous ces problèmes. Par contre, l'analyse de la question doit être en mesure d'aiguiller les personnes responsables du service vers la rédaction d'une réponse satisfaisante à la requête. Le système de question-réponse pour solutionner notre problème doit aller au delà de l'appariement des mots de la question aux mots d'un texte.

3.1.2 Contexte de la question

Le traitement d'une requête dans les services d'information doit considérer l'information provenant du contexte de la requête. Dans le problème que nous étudions, les requêtes proviennent de courriels et ceux-ci contiennent de l'information essentielle à la désambiguïsation de la question. La formulation de la question est elle aussi différente des questions factuelles habituellement traitées par les systèmes de question-réponse. Les questions sont formulées comme des requêtes mais lorsqu'elles sont mises hors-contexte elles n'ont pas le même sens interrogatif.

L'information recherchée, exprimée par la question, est aussi déterminée par le type d'utilisateur qui élabore la question. Pour une même question, l'information demandée par deux utilisateurs aux profils différents sera elle aussi différente. Le système de question-réponse doit considérer le profil de l'utilisateur pour évaluer le besoin d'information d'une question. Si l'utilisateur a déjà utilisé le service, l'information provenant des transactions précédentes doit être utilisée pour désambiguïser la requête.

Le dialogue avec l'utilisateur est une méthode pour évaluer et désambiguïser le besoin d'information, mais il est difficilement réalisable. Lorsqu'on demande à un utilisateur de préciser son besoin d'information il ne répond que très rarement, sauf lorsque l'utilisateur est un « professionnel » de l'information. Dans ce cas, cet utilisateur préfère avoir la possibilité de préciser sa requête dans le sens qu'il veut, en ayant accès à toutes les possibilités de requêtes et d'analyse du système. Ainsi, un système doit être assez flexible pour désambiguïser les requêtes *ordinaires* et offrir une interface à ses outils d'extraction et d'analyse d'information, qui pourront être utilisés par les responsables du service pour accélérer le traitement semi-automatisé des requêtes.

3.1.3 Précision de la réponse

Il est essentiel que le système de question-réponse soit en mesure de déterminer avec quelle certitude la réponse trouvée correspond à ce que l'utilisateur a énoncé comme besoin parce qu'un service d'information n'a pas le droit à l'erreur, il ne doit pas donner de fausses réponses. Pour s'assurer que la réponse à la question soit bonne et pour être redevable de ses décisions, le système doit retourner une justification accompagnant la réponse. Cette justification est utilisée pour expliquer à l'utilisateur le raisonnement réalisé par la machine, comme il est suggéré dans la méthodologie des services de références. L'utilisateur qui n'a pas réussi à bien exprimer son besoin d'information pourra alors s'aider de l'explication de la réponse pour explorer de nouvelles pistes de réponse à sa question ou pour formuler une question plus précise.

Dans le cas d'un système complètement automatisé, la précision du système devra être exemplaire. Mais puisque nous privilégions une approche où le système sera utilisé dans un contexte de support au traitement des courriels, nous avons développé les modules de l'architecture de façon à privilégier le rappel plutôt que la précision des traitements.

3.2 Description de l'architecture

L'architecture que nous proposons (fig. 3.1) est calquée sur les architectures rencontrées en question-réponse factuelle où la réponse est extraite de données textuelles

(fig. 2.1). Le flot de traitement adopté est le même que celui de la question-réponse factuelle. Par contre, la nature des courriels fait en sorte que chaque étape de traitement doit être adapté au problème. Pour solutionner le problème nous devons :

1. identifier la question dans le courriel lors de *l'analyse de la question* (sec. 3.2.1) ;
2. classifier les questions pour sélectionner le traitement à réaliser lors de la *sélection des documents candidats* (sec. 3.2.2) ;
3. extraire les relations lors de *l'analyse des documents candidats* (sec. 3.2.3) ;
4. sélectionner la bonne réponse lors de *l'extraction de la réponse* (sec. 3.2.4) ;
5. rédiger une réponse lors de la *génération de la réponse* (sec. 3.2.5).

Nous n'avons pas inclus la tâche de *prétraitement des documents de références* que nous rencontrons habituellement dans les systèmes de question-réponse dans le flot de traitement puisque ce n'est pas à proprement parler une étape de traitement du courriel. Nous considérons quand même cette étape parce qu'elle peut améliorer la qualité des réponses aux courriels. Le prétraitement des documents nous permet d'identifier les ressources pertinentes du domaine et de construire les bases de connaissances nécessaires pour récupérer les réponses.

Puisque notre architecture se distingue principalement par les traitements qui sont réalisés dans chaque module et par la nature des données que nous avons à traiter dans ce problème, nous décrivons sommairement chaque étape de traitement de notre système dans le reste de cette section. Pour chaque description, nous justifions brièvement le point de vue que nous avons privilégié en fonction des approches traditionnelles de la question-réponse.

3.2.1 Analyse de la question

Le traitement efficace de la question-réponse débute par une analyse du besoin d'information exprimé par l'utilisateur. Pour la question-réponse factuelle, ce besoin d'information est toujours un fait. Dans notre problème, nous devons travailler avec une multitude de besoins d'informations, que ce soit des directives à suivre, des méthodes de calculs, des interprétations, des opinions, etc. Le besoin d'information ne se définit pas en une seule phrase, comme c'est le cas en question-réponse factuelle, nous devons extraire ce besoin du courriel.

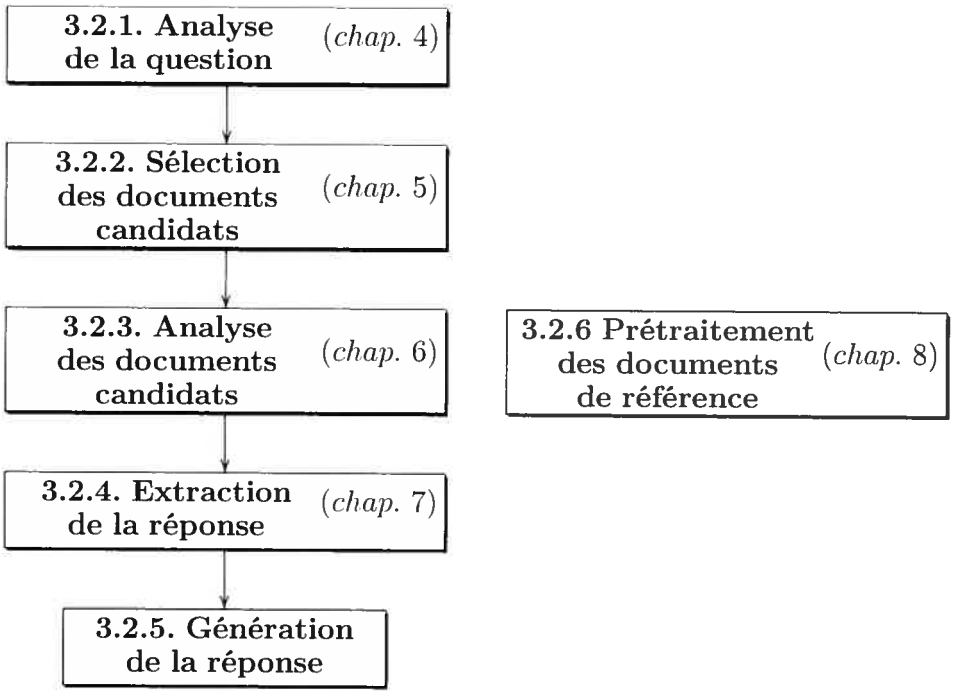

FIG. 3.1 – Architecture question-réponse pour la réponse automatisée aux courriels

La compréhension de la question est une étape essentielle au succès d'un système réponse automatisée aux courriels, comme elle l'est dans la question-réponse. Nous utilisons l'analyse de la question lors de plusieurs étapes de traitement dans notre problème : la sélection des documents candidats, l'extraction de la réponse et la génération de la réponse. Nous avons porté une grande attention au traitement des courriels et des questions, parce que la nature de notre problème est la réponse automatisée aux courriels et qu'une identification erronée de l'information recherchée par l'utilisateur lui donnera assurément une réponse insatisfaisante.

Nous présentons, dans cette thèse, plusieurs travaux qui pourraient faire partie de l'étape d'analyse de la question, mais nous jugeons plus approprié de les présenter dans le contexte où ils seront utilisés. Pour la réalisation d'un système de réponse automatisé aux courriels, nous considérons essentiel d'étudier les données avec lesquelles nous travaillons. Nous débutons le chapitre 4 en présentant le corpus de courriels que nous avons utilisé dans ce travail et qui nous permet d'identifier les problèmes qui doivent être solutionnés lors des étapes subséquentes. Dans ce même chapitre, nous utilisons les résultats de l'analyse du corpus et expliquons comment nous réussissons à extraire les questions des courriels à partir d'une approche à base de règles [6].

3.2.2 Sélection des documents candidats

La sélection des documents candidats dans le système que nous proposons se distingue de ce qui est réalisé en question-réponse factuelle parce que nous n'utilisons pas d'engin de recherche. Puisque dans les courriels nous devons traiter plusieurs types de demandes, nous proposons d'utiliser cette étape pour sélectionner le type de traitement à effectuer pour trouver la réponse. Pour répondre correctement à une requête, nous devons au moins identifier une source et un type d'information ayant un lien avec l'information recherchée.

Le but de la sélection des documents candidats dans un système de réponse automatisé aux courriels est le même que dans la question-réponse factuelle : limiter la quantité de données à analyser précisément. La formulation de la question sous la forme d'une requête à un engin de recherche est une manière de filtrer les sources d'informations pour diminuer la quantité de données à analyser, mais elle ne considère que les caractéristiques de surface de la question. Dans notre problème, nous devons considérer plusieurs types de questions nécessitant une analyse linguistique plus sophistiquée que celle d'une question factuelle. Notre architecture traite la sélection des documents candidats comme une étape de classification de la question. Mais cette classification ne se fait pas d'une seule façon, les multiples caractéristiques d'une question nous permettent de les classifier de façon à restreindre les manières de les traiter.

Dans le chapitre 5, nous étudions le lien entre les caractéristiques de la question et les manières d'interpréter les questions pour y répondre. Pour ce faire, nous utilisons des taxonomies de questions liées aux caractéristiques de la question qui influencent le traitement de la question et la recherche de la réponse. Nous pouvons associer les taxonomies de questions à un niveau d'analyse linguistique pour évaluer la difficulté de la tâche et déterminer la pertinence de l'analyse. Les taxonomies permettent d'identifier les aspects syntaxiques, sémantiques, discursifs et pragmatiques de la question, qui sont pertinents lorsque nous choisissons la méthode pour satisfaire la requête de l'utilisateur. Outre la classification des questions, la représentation formelle de la question nous permet aussi de déterminer le type de traitement que nous utilisons pour répondre à la question.

3.2.3 Analyse des documents candidats

À partir du moment où nous avons identifié les sources d'informations à consulter pour répondre à une question, nous devons analyser ces sources pour identifier l'information essentielle à l'extraction de la réponse. Nous n'utilisons pas directement les données textuelles que nous possédons pour répondre aux questions parce qu'il n'y a que quelques requêtes du corpus qui ont une réponse dans le corpus. L'extraction de la réponse à partir d'une modélisation du domaine est une approche plus pertinente pour notre problème que l'extraction de l'information à partir de données textuelles.

Même si l'utilisation de données structurée est plus pertinente pour solutionner notre problème, l'analyse des données textuelles est essentielle pour utiliser notre solution dans le modèle général de la réponse automatisée aux courriels. L'approche que nous proposons pour analyser les documents consiste à extraire les structures prédicat-arguments. Ces structures peuvent alors être combinées à une ontologie du domaine pour identifier les relations définies textuellement entre les entités. De plus, nous pouvons utiliser les structures prédicat-arguments pour identifier l'information contextuelle contenue dans les courriels et analyser en détail la question en fonction des prédicats utilisés pour exprimer l'information recherchée.

Dans le chapitre 6, nous présentons le repérage des rôles sémantiques comme l'étape d'analyse des documents candidats. Nous débutons par définir les structures prédicat-arguments qui sont utilisées pour représenter les rôles sémantiques. Ensuite, nous définissons le problème du repérage des rôles sémantiques et expliquons comment il est apparut dans notre problème. Pour terminer, nous décrivons les deux approches que nous avons tentées (à base de règles et statistique) et expliquons comment nous avons réalisé le repérage des rôles sémantiques par une méthode d'apprentissage statistique pour analyser les courriels et les documents candidats.

3.2.4 Extraction de la réponse

Lors de l'élaboration de l'architecture, nous avons rencontré un problème d'accès à l'information : nous n'avions pas accès à l'information interne de la compagnie. Nous sommes donc au même niveau que l'utilisateur externe qui soumet sa requête au service d'information. S'il a tenté sans succès d'avoir une réponse à son interrogation

à partir des sources disponibles publiquement, il y a de fortes chances qu'il nous soit aussi impossible d'y répondre. Pour pallier à ce problème, nous avons élaboré une approche pour structurer l'information dans le but de l'interroger.

Nous avons construit une représentation du domaine sous la forme d'une ontologie pouvant ensuite être peuplée à partir des données structurées ou des données textuelles extraites des structures prédicat-arguments obtenues lors d'une étape de prétraitement des documents de référence. L'extraction de la réponse peut alors se faire par une requête à une structure d'information adaptée aux besoins du domaine. La structure d'information doit, prendre en charge les pratiques de l'entreprise au niveau de la gestion de l'information et identifier les modifications de la base de connaissances activées par les événements. En plus de récupérer les réponses, nous devons aussi expliquer comment nous y sommes arrivés, car les utilisateurs veulent parfois avoir les références des sources d'informations pour identifier eux-mêmes l'information qu'ils y cherchent ou pour répondre à d'autres besoins d'informations connexes. Nous présentons les travaux pour modéliser le domaine dans le chapitre 7.

3.2.5 Génération de la réponse

Lors de l'implémentation d'un système de question-réponse dans le cadre d'un service d'information, la génération de la réponse doit être adaptée à l'information retournée et à l'utilisateur. Nous n'avons pas abordé le sujet de la génération de la réponse dans nos travaux. Pour être efficace, la génération automatique de la réponse doit pouvoir compter sur un système retournant toujours la ou les bonnes réponses à un courriel, ce qui n'est pas encore le cas. Dans cette optique, l'architecture que nous proposons devient un système d'aide à la réponse qui, une fois combiné aux travaux de Luc Lamontagne [40], constitue un système de réponse automatisé aux courriels basé sur l'architecture question-réponse.

3.2.6 Prétraitement des documents de référence

Le prétraitement des documents de référence est l'occasion d'appliquer les méthodes d'extractions nécessaires pour identifier l'information essentielle du corpus de référence et remplir les bases de connaissances. L'aspect que nous considérons dans

cette thèse est la temporalité des événements présents dans les articles de journaux, les communiqués de presse et dans les actions d'une entreprise. Notre étude de l'aspect temporel se fait par l'entremise du langage d'annotation du temps et des événements TimeML. Avec ce langage, nous annotons les textes manuellement ce qui nous permet d'identifier facilement les objets temporels et de les mettre en relation. Ce type de prétraitement est applicable à d'autres problèmes que la question-réponse, tel que la génération de résumé.

Dans le chapitre 8, nous présentons notre contribution à l'annotation temporelle avec le langage TimeML : la représentation graphique des objets et relations temporels. De plus, nous y décrivons sommairement le langage d'annotation TimeML servant à étiqueter les marqueurs temporels, les événements et les relations temporelles. L'interprétation graphique et sa représentation est à la base des interactions dans l'outil d'annotation semi-automatique TANGO pour produire des documents annotés avec TimeML.

Chapitre 4

Traitement des courriels

Le choix d'utiliser l'architecture question-réponse provient de l'hypothèse que les courriels envoyés à un service d'information s'interprètent comme des questions nécessitant une réponse. Les données dont nous disposons pour étudier notre problème sont des courriels provenant d'un service de relation avec les investisseurs. Ces courriels sont habituellement des demandes d'information, ainsi, la première étape de traitement à réaliser est l'identification de la question. L'identification de la question dans un courriel est un problème qui a été relativement peu étudié jusqu'ici, la majorité des travaux de recherche concernant le courriel portent sur le filtrage des pourriels (*spam*) et la reconnaissance des réseaux sociaux. Nous avons donc conçu une méthode pour identifier les requêtes dans les courriels [6], élaborée à partir du corpus de courriel (BCE-4) que nous analysons et présentons dans ce chapitre.

L'analyse d'un corpus de référence est essentielle pour évaluer les problèmes que nous devons résoudre pour répondre automatiquement aux courriels. Dans la première partie du chapitre, nous présentons le corpus BCE-4 que nous utilisons pour concevoir et tester les approches que nous proposons. L'analyse du corpus est l'occasion de caractériser et de quantifier les données du problème, comme les premiers systèmes de question-réponse participant à TREC qui analysaient la question avec des modules de traitement créés manuellement à partir des questions des années précédentes. Nous profitons aussi de cette étape pour éliminer les données qui ne sont pas pertinentes à notre problème car tous les courriels ne nécessitent pas nécessairement une réponse.

Les courriels envoyés aux services d'information (services d'assistance, de relation avec les investisseurs ou la clientèle) ne contiennent pas toujours des questions. Lorsqu'ils contiennent une question, l'identification de l'information demandée par la question est plus complexe à réaliser que dans le cas de la question-réponse factuelle. Nous considérons le courriel comme une entité suffisante, constituée d'un ensemble d'informations entremêlées à des interrogations. L'identification précise de la question et de l'information contextuelle contenues dans le courriel est nécessaire pour interpréter correctement les questions formulées dans les courriels. La compréhension du sens de la question est fondamentale pour un traitement automatique efficace des courriels.

La méthode d'identification des questions dans les courriels que nous présentons est inspirée de l'analyse du corpus, mais n'est pas conçue spécifiquement pour notre corpus. Nous utilisons un ensemble de règles sur l'information de surface du courriel pour identifier les questions. Nous présentons les règles formant la grammaire pour identifier les questions dans la deuxième partie du chapitre où nous évaluons la méthode sur le corpus BCE-4.

L'approche à base de règles est plus appropriée que les méthodes à base d'apprentissage statistique parce que la quantité de données disponibles dans les corpus n'est pas assez grande pour que les modèles statistiques soient représentatifs. Les corpus de courriels sont difficiles à obtenir parce que les entreprises ne veulent pas prendre le risque de porter atteinte à la vie privée de leurs « clients ». Le seul corpus de courriel disponible pour la recherche est le corpus ENRON [36], mais il n'est pas directement relié à notre domaine. De plus, il n'était pas disponible lorsque nous avons réalisé cette partie du travail.

4.1 Présentation du corpus de courriels

Les données que nous utilisons pour étudier le problème du traitement automatisé des courriels dans les services d'information ne sont pas des questions isolées comme celles utilisées dans les systèmes de question-réponse factuelle. Dans le cadre du projet de réponse automatisée aux courriels du service de relation avec les investisseurs de la compagnie BCE, nous avons rassemblé le corpus de courriels BCE-4. Celui-ci est le dernier corpus de l'ensemble de corpus BCE-{1,2,3,4} à être utilisé dans notre projet.

Les corpus étudiés n'ont pas tous les mêmes caractéristiques, ce qui nous a permis d'étudier des aspects distincts du problème sur plusieurs ensembles de données.

BCE-1 est constitué de 141 messages envoyés entre avril et septembre 2000 et produits à l'aide d'un formulaire de commentaires accessible à partir du site web de BCE. Ce corpus est composé de messages de nature générale et variée concernant : le site web, les contacts, les relations avec les investisseurs et les plaintes sur la qualité du service. Il a permis de faire une analyse préliminaire du problème de réponse automatisée et du domaine des messages reçus. BCE-2 a été reçu en format imprimé, donc pratiquement impossible à analyser avec des outils informatiques sans avoir préalablement numérisé les documents. Ce corpus est constitué de 865 messages avec le suivi réalisé par les préposés de BCE pour chacun d'eux. L'intérêt de BCE-2 est de contenir un plus grand nombre de messages que BCE-1. Il a été utilisé pour vérifier les hypothèses émises lors de l'analyse de BCE-1 et pour élaborer la proposition de recherche du projet de réponse automatisée aux courriels [37].

BCE-3 contient 1568 paires de message et suivi envoyés à `investor.relations@bce.ca` entre juin 1999 et novembre 2000. C'est le premier corpus qui est représentatif du domaine des relations aux investisseurs, Julien Dubois [20] et Luc Lamontagne [40] l'ont étudié en profondeur dans leurs travaux respectifs. L'étude du corpus révèle que les messages sont difficiles à classifier en utilisant seulement les mots contenus dans les courriels comme attributs dans la classification.

BCE-4 est composé d'environ 1200 courriels envoyés à `investor.relations@bce.ca` et à `relations.investisseurs@bce.ca`, rédigés en anglais et en français. Nous avons réalisé l'étude initiale du corpus à partir de tous les courriels, en considérant autant les courriels rédigés en français que ceux rédigés en anglais. Nous avons choisi de ne traiter que les courriels rédigés en anglais parce que 80% des courriels du corpus sont dans cette langue. Malheureusement, nous n'avions pas assez de données pour traiter efficacement l'utilisation du français dans les courriels des relations aux investisseurs de BCE.

4.1.1 Classification manuelle des courriels

Après avoir analysé le corpus BCE-4, nous avons identifié les courriels contenant des questions pour les classifier selon différentes facettes. Nous avons écarté, pour

les besoins de ce travail, les courriels ne nécessitant aucune action de la part du répondant ($\approx$ 300) et les pourriels ($\approx$ 450), réduisant le nombre de courriels dans le corpus à 236. Nous avons ensuite classifié manuellement les courriels en fonction du sujet de la question et du type de réponse attendu, puis nous avons évalué, pour chaque courriel, s'il pouvait être répondu automatiquement. Cette évaluation de la capacité de répondre à un courriel est très subjective et difficile à mesurer. Notre évaluation de cette capacité est basée sur : l'identification de la question dans le courriel, la présence de l'information complémentaire à la question pour y répondre et l'existence et l'accessibilité de l'information. Le nombre de courriels pour chaque catégorie et le nombre de courriels pouvant être répondus pour chaque catégorie sont présentés dans le tableau 4.1.

Catégories	Nb. courriels	Nb. « répondables »
contact	15	7
dossiers personnels	26	0
date d'événements	24	23
finance et corporations	61	31
comment investir	32	16
prix des actions	59	40
divers	19	9
Total	236	126

TAB. 4.1 – Distribution des courriels du corpus BCE-4

Cette classification des courriels est importante dans ce travail, car elle définit les ressources nécessaires pour répondre aux questions. Les catégories nous permettent de définir les problèmes que nous rencontrons et de lier la nature de ces problèmes au type de requête traitée. Il arrive qu'un courriel contienne plus d'une question, habituellement ces questions peuvent toutes être catégorisées dans une même catégorie. Lorsque les questions d'un courriel peuvent être assignées à des catégories différentes, nous déterminons subjectivement le sujet dominant du courriel et nous catégorisons toutes les questions dans cette catégorie. Puisque nous avons réalisé cette classification manuellement, nous avons considéré le courriel comme une entité indivisible, à la manière d'un préposé qui à la tâche de diriger le courriel vers la ressource la plus compétente pour y répondre. Ainsi, toutes les questions d'un courriel sont considérées comme faisant partie de la même catégorie.

Contact

La catégorie de courriels *contact* est constituée des courriels demandant comment contacter (par courriel ou par téléphone) : une personne dont le nom est mentionné, le responsable d'un département, ou la personne en charge d'exécuter une tâche précisée dans le courriel, ou pour obtenir l'adresse d'une succursale. Les questions de ce type peuvent être répondues automatiquement suite à l'analyse du courriel en utilisant un ensemble de ressources structurées. Le succès de la réponse à ces courriels dépend de la capacité du système à identifier les entités et les relations qui sont utilisées pour récupérer la réponse.

Ci-dessous, nous avons extrait les phrases clés de certains courriels de cette classe. Ces courriels semblent tous posséder une réponse connue, récupérable automatiquement, les questions sont posées clairement et l'information complémentaire nécessaire pour trouver la réponse est contenue dans le courriel. Dans les exemples 1.1 et 1.3, le nom de la personne est clairement indiqué, respectivement John Doe et Foo Bar[1]. Dans l'exemple 1.2, le nom n'est pas mentionné mais la fonction de la personne est clairement indiquée.

1.1. Kindly provide Mr. John Doe's e-mail address. I am a former Vice-President of CBRS ...

1.2. Please advise the name and phone number of your head of Investor Relations.

1.3. Could you please send me an email address and phone number for Foo Bar ?

Certains courriels sont plus difficiles à analyser et l'information pour y répondre n'est peut-être pas accessible par un système automatisé comme le montrent les exemples 2.1, 2.2 et 2.3.

2.1. Please could you let me know if you have any offices in the Caribbean, and if so, where are they and could I possibly have some contact details/emails and addresses ?

2.2. Is there a phone number for the Analyst Meeting on the 12th, instead of watching over the webcast ? Please forward the telephone numbers to listen in to the meeting.

[1]Les exemples présentés ont été modifiés pour éviter de dévoiler des informations personnelles ou des noms de personnes

2.3. We are about to send out a survey via email regarding your DRP/DSPP plan and would like to make sure that it goes to the correct person within your department. ... Can you kindly supply us with the name and email address of this person.

Ces exemples doivent être traités avec attention puisqu'ils contiennent beaucoup d'informations pouvant être interprétées de manière ambiguë. Dans l'exemple 2.1, il n'y a qu'une seule réponse à trouver, mais l'auteur formule sa requête à travers plusieurs questions. L'exemple 2.2 mélange plusieurs moyens de communication, de sorte qu'il est difficile d'identifier l'information recherchée à la première lecture. Pour ce qui est de l'exemple 2.3, il faut découvrir la personne qui a une responsabilité précise dans le département (*your department*). Le traitement des courriels nécessite la reconnaissance de plusieurs types de relations, énoncés de diverses façons, et de types de ressources pour retrouver ces informations.

Dossiers personnels

Nous regroupons les courriels demandant de l'information à propos de cas particuliers dans la catégorie des dossiers personnels. Cette catégorie de courriels contient des requêtes concernant : les cas de successions, la recherche d'actions perdues, la valeur des actions que le client possède et les confirmations d'envois et de réception de papiers. Ces courriels sont les moins susceptibles d'être traités automatiquement, car chaque cas nécessite un traitement individualisé. Les deux principaux facteurs rendant ces courriels difficiles à traiter sont l'imprécision des demandes et la protection de la vie privée. L'imprécision des demandes est attribuable aux auteurs des courriels qui ne donnent pas assez d'informations à propos de leur dossier, parce qu'ils ne connaissent pas assez le domaine de la finance pour poser une question précise. Et nous devons protéger les données des utilisateurs pour ne pas qu'elles aboutissent dans les mains de fraudeurs : les données que nous traitons proviennent de dossiers privés et leur contenu ne doit pas être divulgué sans s'être d'abord assuré de l'identité de la personne. Il nous est impossible de trouver la réponse à ces questions en utilisant les techniques d'extraction d'information utilisées en question-réponse factuelle. Pour ce type de courriel, nous privilégions une approche d'aide à la réponse, à partir d'un ensemble de donnés structurées, et non pas de réponse complètement automatisée.

Les courriels concernant les dossiers personnels proviennent d'investisseurs particuliers, d'exécuteurs testamentaires et de courtiers. Ces courriels mettent en évidence deux catégories d'utilisateurs : l'expert et le novice. L'expert connaît bien le domaine d'application, il exprime sa question clairement et donne assez d'informations pour qu'un préposé (personne ou machine) identifie son besoin d'information et y réponde. Par contre, le novice soumet des requêtes où l'information est plus fréquemment ambiguë ou floue. Quand l'information est imprécise, le système de réponse automatique doit engager une conversation avec l'utilisateur pour qu'il précise sa requête, en posant des questions pour récupérer l'information complémentaire nécessaire à l'identification de l'information recherchée.

Les réponses que nous fournissons aux investisseurs novices doivent être simples. L'auteur de la requête ne distingue pas la précision de la terminologie technique du domaine. Malgré cela, les réponses doivent quand même être complètes pour que le requérant ait au moins l'impression que la réponse satisfait son besoin d'information.

Les courriels provenant des experts, les « *investisseurs éduqués* », contiennent habituellement l'information nécessaire pour que le préposé réponde à leurs requêtes. La difficulté de traitement de ces courriels provient de la nécessité de traiter précisément l'information. Nous ne pouvons pas analyser le message de manière approximative parce que la demande d'information est bien définie et correspond à l'information que l'utilisateur désire recevoir. La précision de la requête n'aide pas à résoudre notre problème, car le jargon du domaine est difficile à comprendre.

Les exemples 3.1 et 3.2 sont représentatifs des courriels concernant les dossiers personnels. L'exemple 3.1 est un cas où l'information est incomplète et la requête impossible à répondre. L'information demandée dans cet exemple nécessite une recherche approfondie à travers plusieurs sources d'informations et l'auteur du courriel ne semble pas être un expert du monde de la finance ; dans ce cas-ci, un dialogue avec l'utilisateur est nécessaire pour qu'il obtienne une réponse satisfaisante à sa question.

L'information de l'exemple 3.2 est complète et la question est clairement formulée. Cet exemple met en évidence le fait que l'analyse du courriel ne doit pas être faite seulement sur le texte où nous retrouvons l'information à propos du certificat en question, mais aussi sur l'information complémentaire, par exemple le sujet du courriel. Dans cet exemple, l'information demandée par l'auteur du message doit être retournée de façon précise et correcte puisqu'il semble connaître le domaine et savoir

ce qu'il recherche. Si l'auteur avait eu accès à l'information qu'il recherche, il n'aurait pas eu besoin de formuler sa requête à un préposé par le biais d'un courriel, il aurait trouvé lui-même l'information recherchée.

3.1. I am the executor for my father's estate.
I have not record that at his death Sept. 6th, of this year he held any BCE stock.
Could you please check your records and determine if any are held at the moment.

3.2. **Subject :** Share Certificate # DC RR928007
...
Can you please advise the history, current status and value of this certificate ?

La formulation des courriels à propos des dossiers personnels est similaire d'un courriel à l'autre. L'acte de parole exprimé par le courriel nécessite une réponse, mais l'acte n'est pas exprimé de façon directe comme c'est le cas dans les questions factuelles. L'ajout de formules de politesse complexifie l'analyse du besoin d'information du message. Puisque ces courriels nécessitent un traitement complexe pour être en mesure d'y répondre une fois la question identifiée, il est préférable qu'un préposé contacte directement la personne pour traiter sa requête.

Dates d'événements

Les courriels concernant les dates d'événements nécessitent des ressources appropriées pour être répondus, telles qu'une chronologie des événements passés et un agenda des événements à venir. Les types d'événements traités par le service de relations aux investisseurs que nous avons établis à partir du corpus sont :

- les dates d'émission et de division d'action ;
- les dates de remise de dividendes ;
- les dates d'émission et d'encaissement des obligations ;
- les dates des différentes conférences où les investisseurs sont invités à prendre part.

La référence au temps dans ces courriels se fait indirectement par l'utilisation d'expressions temporelles relatives comme *next*, *last*, *since*. Les exemples 4.1–4.7 proviennent des messages que nous avons classifiés dans la catégorie des courriels faisant référence au temps ou aux événements corporatifs.

4.1. Could you please let me know the date and location of your 2002 annual meeting. Also, do you have a list of the release dates for your 2002 quarterly results ?

4.2. Would you be able to supply me with the date and location of BCE's next AGM ?

4.3. Could you please advise as to when the rates will be posted ?

4.4. **Subject :** When will BCE be reporting its Q4 2001 results ?

4.5. Would you advise me of the BCE stock splits since February 1994.

4.6. According to our records, we are expecting a dividend announcement from you. Please could you let me know when this announcement is scheduled to happen (approximately), i.e. :

1. The expected ANNOUCEMENT/DECLARATION date for the dividend to be paid in JANUARY 2002.
2. Could you also confirm the PAY and RECORD date for this dividend yet to be announce ?
3. In addition could you also tell me when your next Annual Shareholders Meeting is ?

4.7. Can you please confirm what time the first presentation will begin at for the meeting on December 12th.

L'exemple 4.1 est représentatif des demandes d'informations à propos des événements à venir. Ce type de message est constitué d'une à trois phrases courtes et la demande d'information est formulée de façon très précise. L'exemple 4.2 est similaire, mais il utilise l'abréviation *AGM* pour identifier *Annual General Meeting*. Il est donc important de reconnaître les abréviations et les collocations pour identifier la date et le lieu de l'événement dont l'utilisateur s'enquiert. Nous pouvons remarquer un problème similaire dans l'exemple 4.3, où l'objet de la question est mal défini. Lorsque l'utilisateur mentionne *rates*, il est impossible (même avec le contenu complet du message) de déterminer à quel taux le message fait référence. Et l'interprétation de l'expression *to post the rates* n'est pas la même dans le domaine des investisseurs que dans le domaine des connaissances générales. L'exemple 4.4 démontre, tout comme l'exemple 3.2, que toute l'information disponible doit être utilisée ; dans cet exemple, le corps du message est vide et la question se retrouvant dans le champ *Sujet* est l'unique contenu du courriel. L'extrait de l'exemple 4.5 concernant les *stock splits* est fréquent dans le corpus. Nous avons seulement trois courriels comme celui-ci dans la catégorie *dates d'événements*, mais nous en retrouvons douze autres très

similaires dans la catégorie *prix des actions* qui peuvent être considérés comme une sous-catégorie des événements. Les questions de ce type se formulent différemment, ainsi, certaines peuvent être répondues directement alors que d'autres demandent un processus de raisonnement, comme dans l'exemple 4.5 où un comptage doit être réalisé.

Dans l'exemple 4.6, nous retrouvons trois questions parmi lesquelles nous avons un cas de réutilisation de réponse avec les deux premières questions. Ces deux questions, tout comme la troisième, concernent des événements futurs dont l'information n'existe possiblement pas au moment de la réception du message. Dans ce cas, nous devons établir avec certitude que l'événement est valide, mais que la date n'est pas encore établie et identifier le moment où la date de l'événement sera déterminée pour répondre au courriel. L'exemple 4.7 est très précis, il demande un programme des présentations de la journée lors de la rencontre du 12 décembre.

Les questions des exemples précédents sont généralement bien formulées et leur traitement n'est pas problématique pour un préposé. La recherche de la réponse est un problème plus important que l'analyse de la question pour traiter automatiquement les courriels concernant les événements. L'information temporelle est difficile à extraire automatiquement à partir des textes et les événements futurs ne sont pas nécessairement dans les bases de connaissances.

Finance et corporation

La catégorie finance et corporation inclut les messages concernant les finances, le fonctionnement et les publications de BCE (p. ex. rapport annuel). Nous avons ajouté à cette catégorie tous les courriels concernant : la structure de BCE, le nombre d'employés, la répartition des investisseurs et l'information commerciale à propos des produits de BCE. La réponse automatisée à ces questions est délicate car certaines informations peuvent être confidentielles ou inaccessibles, parce que la compagnie ne veut pas les divulguer ; ce qui explique que l'information soit demandée par l'entremise d'un courriel suite à une recherche infructueuse de la part de l'utilisateur.

Les courriels concernant la finance sont difficiles à comprendre, même pour les personnes habilitées. Nous ne devons donc pas avoir des attentes trop élevées envers un système automatique pour répondre à ce type de questions. Les problèmes que

nous rencontrons sont la technicité du langage et l'interprétation de la terminologie faite par les intervenants. L'information utilisée pour répondre à ce type de requêtes est variable dans le temps ; comme dans l'exemple 5.1, où le pourcentage d'actions détenues par BCE à titre d'actionnaire dans Nortel Networks diffère en fonction de la date à laquelle l'information est demandée.

5.1. Please tell me if BCE Inc. still owns any shares of Nortel Networks Corp., and if so, what percentage of Nortel's outstanding shares does BCE Inc.'s holdings represent.

5.2. Hi, I am a fourth year student at Wilfrid Laurier University and have an assignment about BCE for my Investments Management class.
I am hoping that you will be able to give me the following information.
The P/E Ratio for 1995-2000.
The average stock price for each year 1995-2000
The market and book value of the common shares.

5.3. Would you please be able to let me know what were your EBITDA for 1998, 1999 and 2000 and how they were calculated (actual earnings, interests, taxes, depreciation and amortization figures).

L'exemple 5.2 contient trois questions et chacune d'elle nécessite plus d'une réponse. La réponse aux questions contenant des intervalles de temps comme 1995–2000 doit être traitées avec attention parce qu'elle doit contenir l'information pour l'ensemble des années demandées. En analysant chacune des questions, nous identifions certains problèmes à résoudre pour réaliser l'analyse automatique.

La première requête de l'exemple 5.2 s'interprète de deux façons, la personne veut avoir le *price per earnings ratio* (*P/E Ratio*) pour l'ensemble de la période 1995–2000 ou la personne désire le *price per earnings ratio* pour chaque année de 1995 à 2000. Cette dernière interprétation est influencée par le contexte qui, dans la deuxième question, spécifie explicitement qu'il veut une valeur pour chaque année, par l'utilisation de *each year*. La troisième question s'interprète de la même façon que la première sauf qu'elle ne mentionne pas spécifiquement la période de temps.

L'exemple 5.3 peut être répondu, malgré une syntaxe difficile à analyser, car l'information contenue dans la question est complète. La première partie de l'exemple demande l'EBIDTA[2] pour certaines années. Dans la deuxième partie, nous devons utiliser un raisonnement plus complexe où il faut savoir comment sont calculées les

[2] « Earnings before interest, taxes, depreciation and amortization, [...] The term "EBITDA" does not have a standardized meaning prescribed by Canadian generally accepted accounting principles. » (source BCE)

différentes données financières. Ce type de question pose des problèmes à deux niveaux.

Le premier problème se situe au niveau de l'extraction des connaissances, l'identification de cette information à partir de données textuelles est une tâche qui demande une précision parfaite de la part d'un engin de forage d'information.

Le second problème se situe au niveau de la représentation des connaissances, comment devons-nous encoder les méthodes de calculs ? Nous pouvons utiliser des formules mathématiques et décrire de façon explicite chaque variable de façon à avoir un modèle complet du système financier. Nous pouvons aussi utiliser des noms de méthodes auxquelles nous ajoutons une description. En réalité nous devons faire des compromis pour obtenir une représentation fusionnant les deux formes de représentation pour les utiliser dans la recherche de la réponse.

Comment investir

Dans cette catégorie, nous considérons les courriels des investisseurs (ou futurs investisseurs) demandant comment investir dans les différents produits financiers proposés par BCE, auxquels nous ajoutons les courriels de ceux qui veulent savoir comment réinvestir leurs dividendes. Une fois les questions extraites des courriels, un préposé est capable de répondre à ces questions en consultant le site web de BCE. Dans certains cas, la question formulée dans un courriel est similaire à la description de certains documents mis à la disposition des investisseurs par BCE. Les réponses à ce type de courriels sont longues et ne peuvent pas être une expression ou une phrase courte. Ces questions sont sujettes à subir une deuxième étape de classification selon la forme de la réponse attendue. Les formes de réponses les plus communes pour cette catégorie sont : des directions, des descriptions, des instructions et des citations.

Les courriels de cette catégorie sont les plus difficiles à traiter, si on exclut les courriels de la catégorie *dossiers personnels*. Les difficultés résident dans le fait que nous devons déterminer avec certitude la bonne réponse et la réponse doit être rédigée avec attention pour ne pas introduire de propos qui pourraient être ambigus. Les réponses à cette catégorie de questions influencent les décisions des investisseurs et peuvent avoir de graves répercussions sur celui-ci s'il reçoit une information erronée ou une explication imprécise ou inappropriée. De plus, même si le courriel peut être

analysé et répondu, suite à la réponse, l'investisseur voudra assurément discuter avec un préposé ou avec un courtier. Un autre facteur qui influence la réponse est le lieu de résidence de l'investisseur, une facette du problème qui nous était inconnue avant d'examiner certains documents provenant du site web de BCE. Les procédures d'investissements au Canada, aux États-Unis et ailleurs dans le monde ont toutes leurs particularités, ce qui complique la problématique d'extraction de la réponse et fait en sorte qu'une personne doit toujours faire le suivi de la requête. Nous rencontrons aussi des utilisateurs qui veulent être rassurés quant à leur démarche, pour eux le traitement automatisé de leur requête n'est pas approprié.

Prix des actions

Le prix des actions est l'information la plus demandée, ce sont principalement des actionnaires ou ex-actionnaires qui demandent le prix d'actions à différentes dates. Parmi ces courriels, certains proviennent d'investisseurs demandant comment calculer la formule d'allocation reliée au « spin-off » de Nortel, certains demandent comment calculer le coût de base ajusté (pour faire leur rapport d'impôt) et d'autres contiennent des questions reliées au calcul du prix des actions. Les difficultés que nous rencontrons dans ces courriels sont au niveau de l'extraction de l'information contenue dans le courriel. Si cette information est complète et extraite convenablement, la tâche est alors d'exécuter des calculs bien définis à partir de données existantes sur le site web de BCE, comme les cours de fermeture quotidiens des actions ordinaires de BCE présentés dans un fichier de chiffrier électronique (MS Excel).

Les questions qui nous causent des problèmes sont celles qui font appel à des données qui ne sont pas facilement accessibles ou bien à des calculs spécifiques, comme dans l'exemple 6.1, où le prix moyen des actions (*average trading price*) n'est pas connu et qu'il doit être calculé selon les formules comptables de BCE.

6.1. Could you provide me with the average trading price for BCE for the years 1970 and 1979.

6.2. Please advise the closing price of BCE common shares on April 27, 1998.

6.3. COULD you please let me know what the annual dividend rate has been established at for the Series T Preferred Shares.

L'exemple 6.2 correspond aux questions que nous retrouvons le plus souvent dans cette catégorie, soit une demande du prix de fermeture de l'action de BCE à une date précise. Cette information se retrouve aisément à partir de données tabulaires. L'exemple 6.3 est problématique pour un système de réponse automatisé, puisque les actions privilégiées de série T ont été converties en actions d'une autre série, un phénomène fréquent dans les entreprises publiques. Comme elles ont été converties, elles n'existent plus. Pour répondre à cette question, nous devons donc extraire toute l'information disponible et lier cette information. Dans ce cas-ci, le taux demandé n'est pas disponible à partir du site web, mais avec une connaissance du domaine et les ressources nécessaires, un préposé peut répondre à cette question.

Dans ces exemples, nous devons considérer l'aspect temporel de l'information avec attention. Les dates sont indiquées clairement dans les exemples 6.1 et 6.2, mais l'information temporelle de l'exemple 6.3 est par contre plus difficile à identifier. La temporalité dans cet exemple réfère à la chronologie des événements se rapportant à un titre dont l'état change dans le temps. Si nous traitons le message dans l'intervalle où ce titre est transigé, le problème de l'existence du titre ne se présente pas. Dans le cas contraire, nous déterminons premièrement si l'information que nous retrouvons dans les données est encore pertinente lorsque la question est traitée.

4.2 Identification des questions

Les courriels reçus dans le cadre d'un service d'aide contiennent des questions ou des requêtes et ils nécessitent toujours une réponse de la part d'un préposé. L'identification de la requête peut se faire à partir de patrons créés à partir des formes traditionnelles de questions et d'un corpus de référence, en l'occurrence nous utilisons BCE-4.

Pour identifier la question, nous avons créé une grammaire qui identifie les patrons lexicaux les plus susceptibles d'être des requêtes. Nous avons implémenté la grammaire dans un module de traitement linguistique de l'environnement d'ingénierie linguistique GATE dans le but d'extraire le plus grand nombre de questions. La solution que nous avons conçue vise un taux de rappel des questions élevé au détriment de la précision. Dans le cadre du service à la clientèle, nous considérons qu'il vaut mieux traiter automatiquement quelques « questions » supplémentaires que d'ignorer

(involontairement) une question pertinente. Lors de l'identification des questions, nous n'avons pas utilisé d'analyses sophistiquées de courriels, nous avons conservé le traitement linguistique plus évolué pour traiter la question lorsqu'elle sera identifiée.

4.2.1 Grammaire d'identification des questions

Nous avons implémenté la grammaire d'extraction des questions comme un module de traitement de la langue, intégré au système d'ingénierie linguistique GATE de l'Université Sheffield [17], en utilisant à la fois de l'information syntaxique et lexicale. Nous avons pris cette décision parce que l'environnement GATE nous permet de créer des modules indépendants, mais pouvant être utilisés ensemble dans un flot de traitement. Le procédé de détection des questions s'exécute en une série d'étapes. Nous effectuons les premières étapes de traitement à partir d'une suite de modules de traitement linguistique usuels provenant du système GATE. Nous détectons les questions dans le module d'identification des questions qui est exécuté en deux étapes :

1. l'identification, par un *gazetteer*, des mots dans les courriels qui sont utilisés comme symboles terminaux dans la grammaire ;
2. La deuxième étape du traitement est l'identification de patrons de questions, encodée comme une cascade de transducteurs et exprimée dans le formalisme JAPE.

Nous pouvons examiner les deux composantes du module présentées sous la forme synthétisée d'une grammaire dans la figure 4.1. Nous numérotons chacune des règles d'identification et écrivons en *italique* le nom de la règle, les symboles écrits en majuscules et en *italiques* sont des symboles non-terminaux et les symboles écrits avec la police `courrier` sont des terminaux.

La tâche de chaque règle est d'identifier un patron correspondant à une façon de poser une question dans un courriel. L'application des règles consiste à faire une analyse en surface du courriel. La question identifiée correspond à la région du courriel commençant par le début du patron et se terminant au symbole de fin de phrase déterminé par le module de segmentation de phrase. Nous définissons les règles de la figure 4.1 de la façon suivante :

(4.1) Les patrons `I was wondering` ou `I wonder` de la règle *wonder* sont utilisés dans les courriels pour introduire poliment une question ou une requête formulée

wonder → I (wonder | was wondering) (4.1)

be_have_mod → (*MODALS* | *TOBE* | *TOHAVE*) (4.2)
(it | there | these | *EX*)[*DET*]

modalsBegin → (*MODALS* | *TODO*)(*PRP* | *NNP*) (4.3)

actionAsking → *ACTION_ASKING* (me | us) (4.4)

wh_be_have_do → *WH_WORDS* (*TOBE* | *TOHAVE* | *TODO*) (4.5)

Please → please *ACTION_ASKING* (4.6)
category=VB

would_like_to → [I | we | he | she | and] would like to (4.7)
category=PRP

wh → *WH_WORDS* [category = PRP] (4.8)

MODALS → can | could | may | might | must | shall |
should | will | would | ought to

ACTION_ASKING → advise | forward | provide | confirm |
give | send | direct | let | tell

WH_WORDS → what | where | when | which | who |
why | how

FIG. 4.1 – Grammaire d'identification des patrons

indirectement. L'information recherchée pour identifier précisément la requête ne suit pas immédiatement le patron recherché, le patron « if you *modals action* » se retrouve régulièrement entre le patron recherché par la règle et l'information recherchée.

(4.2) La règle *be_have_mod* identifie des questions portant sur la confirmation d'une information incomplète ou sur la vérification de l'existence d'une entité (compagnie, méthode de calcul, produit financier, ...) Ces questions sont compliquées à analyser parce qu'il faut déterminer l'information cherchée par l'auteur. La création de la réponse est encore plus complexe puisqu'elle dépend directement de la confirmation de la requête.

(4.3) Les questions identifiées par la règle *modalsBegin* sont difficiles à catégoriser, le patron que nous utilisons correspond à une syntaxe peu spécifique pour introduire une question. Cette règle peut être considérée comme un cas général de la règle *wonder*. Suite à l'identification de la question, les réponses à chercher sont déterminées par l'action indiquée par le verbe suivant le patron.

(4.4) La règle *actionAsking* identifie les questions où le préposé doit exécuter une action. Cette règle sert aussi pour certaines formulations de requêtes similaires aux règles précédentes, mais où la partie d'introduction de la requête est absente ou pas assez fréquente pour être considérée. Les verbes d'actions utilisés pour identifier les requêtes sont : `advise`, `forward`, `provide`, `confirm`, `give`, `send`, `direct`, `let` et `tell`.

(4.5) La règle *wh_be_have_do* identifie les questions qui débutent par un *wh-words* suivi d'une forme des verbes *be*, *have* ou *do*. Ces questions sont parmi les plus communes, leur forme est généralement celle à laquelle on fait référence lorsqu'on considère le domaine des questions.

(4.6) La règle *Please* est très similaire à la règle *actionAsking*, la composante importante pour les deux règles est le verbe énonçant l'action demandée. La différence avec la règle *actionAsking* est que le verbe d'action n'est pas nécessairement utilisé de façon transitive.

(4.7) Les requêtes identifiées par la règle *would_like_to* sont des questions où le but n'apparaît pas clairement. Le but de la question dépend du verbe suivant `to` dans le patron, celui-ci peut être *ask*, *confirm*, *inquire*, *know*, *attend*, ...

(4.8) La règle *wh* est réalisée pour récupérer l'ensemble des questions contenant un *WH_WORD* et qui n'ont pas été identifiées auparavant. Les questions identifiées n'ont pas de caractéristiques particulières, elles devront être traitées avec des typologies de questions similaires à ce qui se fait dans systèmes de question-réponse pour pouvoir être répondues.

L'identification des questions se fait en essayant les règles à tour de rôle pour identifier celle dont le patron concorde avec le texte. Lorsqu'un patron de règle concorde, l'intervalle de texte débutant au premier mot du patron et se terminant à la fin de la phrase est identifié comme une question. Le mot suivant la fin de la question est ensuite utilisé pour débuter la recherche de la question suivante. Cette stratégie de recherche nous permet d'appliquer un ordre de priorité sur les règles. Dans le module

d'identification des questions, nous avons établi l'ordre de priorité des règles selon leur ordre d'apparition dans la grammaire.

4.2.2 Présentation des résultats

Nous avons tout d'abord vérifié sur le corpus BCE-4 que les courriels contenant des questions ou des requêtes étaient identifiés par notre module. Dans le tableau 4.2, nous considérons qu'un courriel est bien identifié lorsqu'au moins une question est identifiée correctement dans le courriel. Un courriel est mal identifié lorsque la ou les questions qu'il contient ne sont pas ou ont été mal identifiées. Nous présentons les résultats de l'identification des questions (tab. 4.2) en fonction des catégories de courriels que nous avions préalablement classifiés manuellement (sec. 4.1.1). Nous ne considérons pas les courriels de la catégorie des dossiers personnels parce qu'ils sont très difficiles à répondre, même par une personne, et ils nécessiteraient chacun une règle dans notre grammaire pour identifier les requêtes qu'ils contiennent. La répartition de l'évaluation sur les six catégories de courriels nous donne un meilleur aperçu des difficultés d'identification des questions reliées à chaque type de courriel.

Catégorie	Nbr. courriels	Bien identifiés	Mal identifiés	« Rappel »
contact	15	11	4	0,73
date	24	23	1	0,96
divers	19	17	2	0,89
finance	61	50	11	0,82
invest	32	26	6	0,81
share	59	44	15	0,75
Total	210	171	39	0,81

TAB. 4.2 – Évaluation de l'identification des questions pour chaque catégorie

Dans l'ensemble, l'identification des courriels contenant des questions est satisfaisante, 81% des courriels contenant une ou des questions sont identifiés comme tel (tous les courriels en contiennent au moins une). L'identification est particulièrement efficace pour les courriels des catégories *date* et *divers*, où les courriels sont identifiés à 96% et 89% respectivement. Inversement, l'identification est moins performante pour les catégories *share* et *contact*. La différence de performance entre les catégories est attribuable à la forme plus conventionnelle des questions diverses et de celles

concernant les dates, par rapport au côté plus personnel des questions concernant les actions et les coordonnées.

Nous considérons ensuite l'efficacité de l'identification des questions au niveau des questions contenues dans chaque courriel. Nous présentons, dans le tableau 4.3, la distribution de l'efficacité du module d'identification de questions en fonction des questions identifiées manuellement. Nous avons réalisé l'identification manuelle des questions suite à la conception de la grammaire, au moment de l'évaluation de la tâche. Les données sont maintenant exprimées en fonction de la distribution de la bonne ou mauvaise identification des questions et non plus seulement par rapport à l'identification des courriels. Cette évaluation est importante parce qu'elle tient compte du fait qu'il y a plus d'une question par courriel. Et comme nous le mentionnions précédemment, toutes les questions d'un courriel sont catégorisées dans la même catégorie. Dans notre cas, nous identifions 363 questions parmi les 210 courriels. La plupart des courriels contiennent une seule question, lorsqu'un courriel contient plusieurs questions, elles touchent généralement le même sujet ou des sujets apparentés tels que *finance*, *invest* ou *share*. Nous considérons qu'une question est bien identifiée lorsque l'extrait reconnu correspond assez bien à la question repérée manuellement. Une question mal identifiée chevauche une question identifiée manuellement, mais elle ne peut être utilisée comme donnée d'entrée à un système de question-réponse. En considérant le nombre total de questions bien identifiées par rapport au nombre total de questions, nous obtenons un taux de rappel de 79%. Les questions non identifiées, comparativement aux questions mal identifiées, n'ont aucun segment qui a été identifié comme une question. Ce sont des questions qui doivent être identifiées d'une autre façon ou qui sont formulées sous la forme d'une phrase affirmative.

Suite à l'analyse du rappel de l'identification des questions, nous avons voulu connaître la précision du procédé. Si nous considérons les questions mal identifiées comme des résultats négatifs alors nous identifions 67 questions (identifiées à tort + mal identifiées) qui n'en sont pas, ce qui donne une précision de 81%. Dans l'interprétation des résultats, nous devons considérer que les données utilisées contiennent toutes une question et les questions mal identifiées sont quand même utiles, parce qu'elles indiquent, lorsque le courriel est traité par une personne, la présence d'une question potentielle.

Catégorie	Nombre de questions	Bien identifiées	Mal identifiées	Identifiées à tort			Non identifiées
				normal	sig.	rép.	
contact	16	11	3	1	2	0	2
date	27	27	0	2	3	0	0
divers	42	30	0	2	0	4	12
finance	147	112	10	6	10	2	25
invest	50	41	5	4	1	0	4
share	81	65	4	3	5	0	12
Total	363	286	22	18	21	6	55

TAB. 4.3 – Distribution des identifications par l'extracteur de questions pour les courriels contenant des questions

Nous identifions trois catégories pour classifier les « questions » ayant été identifiées à tort comme des questions. En examinant les erreurs, nous distinguons les catégories d'erreurs la localisation de l'erreur dans le courriel.

1. La première catégorie de questions identifiées à tort est celle où le patron de question apparaît dans le corps principal du courriel, mais où ce patron n'est pas une question. La plupart du temps c'est un *WH_WORD* agissant comme un pronom relatif pour introduire une clause relative, qui n'est pas utilisée dans un contexte interrogatif. L'exemple 7.1 en est un où une question est mal identifiée, l'identification est réalisée par la règle *modalsBegin*.

   7.1. Please do not hesitate to contact me **should you need clarification or have any inquiries.**

2. La deuxième catégorie de questions identifiées à tort est attribuable aux questions extraites dans la partie signature du courriel (identifiée par sig. dans le tableau 4.3.) La phrase extraite doit parfois être considérée comme une question, mais la plupart du temps ce n'est pas le cas, l'identification du patron ne correspond pas à une question ou le patron est une formule toute faite du type Do you Yahoo!?, Where do you want to go today? ou tout autre slogan publicitaire énoncé comme une question. Cette partie du courriel aurait pu être éliminée, mais nous avons choisi de la conserver car elle peut nous donner de l'information sur le type d'utilisateur demandant une information.

3. La troisième catégorie de questions qui doit être ignorée lors de l'identification est celle où la question apparait dans la partie retour (*reply*) d'un courriel (identifiée par rép. dans le tableau 4.3). Ces questions doivent être traitées avec attention, nous ne pouvons pas simplement ignorer le contenu en citation. Quelques fois les questions citées proviennent d'un courriel redirigé qui doit être répondu, tandis qu'à d'autres occasions ce sera seulement une information qui a suivi le fil d'une *conversation* par courriel.

En analysant les résultats en détail, nous constatons que les courriels concernant les questions financières de la compagnie et ceux ayant un lien avec le prix des actions sont plus difficiles à traiter. Nous expliquons cela par la complexité des énoncés de questions et l'utilisation du contexte du courriel pour déterminer le sens interrogatif du message. De plus, la catégorie finance possède la plus grande densité de questions annotées par courriel, soit 112 questions réparties à l'intérieur de 61 courriels. La densité de questions par courriel ($\approx$ 2.35 questions/courriel) complique le traitement automatisé, car nous devons conserver le contexte des questions du courriel pour identifier et répondre efficacement à chaque question.

Dans une troisième étape, nous examinons les résultats de l'identification pour chaque règle. Dans le tableau 4.4, nous indiquons le nombre de questions identifiées par chaque règle pour chaque catégorie de courriels. La présentation des résultats nous donne une distribution des patrons de questions qui n'est pas uniforme. Nous devons noter, pour le lecteur attentif, que le nombre de questions identifiées par les patrons est plus grand que le nombre de questions « identifiées » du tableau 4.3 (p. 60). Ceci est simplement attribuable au fait qu'une question peut être identifiée à tort ou à raison par deux patrons, comme dans le cas où une abréviation est considérée comme une fin de phrase, un segment correspondant à une question peut débuter immédiatement après.

De ces résultats, nous remarquons que :

- 270 des 373 questions annotées du corpus l'ont été avec seulement deux règles (*modalsBegin* et *wh-words*) ;
- les règles *wonder* et *actionAsking* ne sont utilisées que pour les catégories de courriels *finance* et *share* ;
- les règles *Please*, *would_like_to* et *be_have_mod* semblent être activées de façon uniforme relativement à la quantité de questions des catégories de courriels ;

Catégorie	*wh-words*	*modalsBegin*	*Please*	*actionAsking*	*would_like_to*	*be_have_mod*	*wonder*	*wh_be_have_do*	Total
contact	4	7	2	1 sig.	3	2	0	0	19
date	8	22	5	1 sig.	2	0	0	0	38
divers	9	15	3	0	0	7	0	0	34
finance	87	32	4	13	1	4	0	0	141
invest	15	22	6	0	5	3	6	0	57
share	18	31	15	9	7	2	2	0	84
Total	141	129	35	24	18	18	8	0	373

TAB. 4.4 – Distribution des patrons de questions retrouvées dans chacune des catégories

– la règle *wh_be_have_do* n'identifie aucun patron dans les courriels.

Nous avons ensuite modifié la méthode d'application des règles de la grammaire pour permettre à toutes les règles d'être appliquées même si le segment de texte a déjà été identifié. Nous avons aussi modifié le module JAPE de l'environnement GATE car certaines fonctionnalités avaient des comportements étranges et les résultats de l'identification de la question pouvaient varier selon le système d'exploitation utilisé. En autorisant l'application de toutes les règles de grammaire et en utilisant le module de transducteur JAPE modifié, nous avons solutionné le problème de stabilité des résultats. Suite à une analyse superficielle des résultats, nous avons remarqué que nous reconnaissions un plus grand nombre d'extraits. Cette augmentation est du à l'identification des chevauchements de segments de texte (tab. 4.5). Malgré l'augmentation du nombre d'extraits identifiés, cette façon de procéder n'améliore pas l'identification de la question parce que nous récupérons plus de questions identifiées à tort que de bonnes questions. L'identification d'une question par plusieurs règles nuit à la précision de l'identification puisque nous devons déterminer quel segment de texte, parmi ceux retournés par les règle, représente le plus précisément la requête.

Catégorie	*wh-words*	*modalsBegin*	*Please*	*actionAsking*	*would_like_to*	*be_have_mod*	*wonder*	*wh_be_have_do*	Total
contact	5	8	2	1	4	2	0	0	22
date	6	32	7	2	0	0	0	0	47
divers	15	18	5	0	3	10	0	0	51
finance	97	37	4	19	1	5	8	0	171
invest	20	31	8	0	5	4	0	0	68
share	21	42	19	12	12	2	2	0	110
Total	164	168	45	34	25	23	10	0	469

TAB. 4.5 – Distribution des patrons de questions retrouvées dans chacune des catégories (avec possibilité de chevauchement).

4.2.3 Discussion des résultats

Les résultats que nous obtenons avec notre grammaire d'identification des questions confirment qu'à partir d'un ensemble limité de règles générales nous couvrons un vaste ensemble de questions. Mais la grammaire ne couvre pas toutes les formes de questions qui se retrouvent dans le corpus, les résultats (tab. 4.3) indiquent que la grammaire n'a pas été en mesure d'identifier 55 questions (15% du nombre total de questions).

Le nombre de questions non-identifiées n'est pas une contre-performance de la méthode d'identification des questions, car les erreurs s'expliquent par des exemples pathologiques du corpus. Cinq courriels dans le corpus contiennent près de la moitié des questions non-identifiées (26 questions). Ces cinq courriels ne sont pas représentatifs des données du corpus, ils ont une densité de questions très élevée et ils utilisent le formatage du texte et non la syntaxe pour distinguer les questions. Les autres questions non-identifiées sont dues à deux facteurs : le premier est la faible densité des patrons utilisés pour les questions non-identifiées et le deuxième est la présence de structures grammaticales incorrectes, de coquilles et de phrases difficiles à comprendre.

L'approche par patron de surface est la plus adaptée à notre problème. Nous ne pouvons utiliser des méthodes d'apprentissage pour identifier les questions puisque la quantité de données n'est pas assez grande et le corpus de courriels n'est pas assez diversifié pour trouver des facteurs discriminants assez forts. Nous pouvons utiliser les règles dans n'importe quel service d'information pour identifier les questions parce que les résultats que nous obtenons sont très bons, mais surtout la méthode n'utilise pas de caractéristiques des courriels propres au domaine des investisseurs. Nous pouvons améliorer l'identification en réalisant un traitement linguistique plus évolué, mais à cette étape du problème il est prématuré de le faire. Lorsque nous appliquons cette méthode d'identification, nous devons toujours avoir en tête la relation entre la précision et le rappel. Nous avons tenté d'améliorer la grammaire en ajoutant des règles et en spécialisant les règles existantes, mais à chaque fois qu'une mesure (précision ou rappel) augmente, le gain réalisé est annulé par une diminution plus grande de l'autre mesure.

4.3 Conclusion

Dans ce chapitre nous avons présenté le corpus de courriel BCE-4, utilisé pour valider les pistes explorées pour les étapes de traitement. De cette analyse, nous avons mis en évidence que nous traitons un problème difficile, plus compliqué que celui de la question-réponse factuelle parce que nous devons considérer plusieurs sources d'informations et que les utilisateurs ne cherchent pas seulement l'information mais aussi la façon de l'obtenir.

La classification manuelle des courriels est importante parce qu'elle nous permet d'identifier les problèmes spécifiques à chaque type de courriel. Cette classification ne suit pas un critère en particulier puisque nous avons classifié les courriels comme un préposé l'aurait fait. Rétrospectivement, nous pouvons constater que chaque catégorie correspond à une combinaison d'aspects et non pas à un seul. Nous pourrons voir dans le prochain chapitre que les catégories que nous avons définies correspondent à plusieurs aspects de la classification qui sont issus des systèmes de référence numérique.

Dans un deuxième temps, nous avons pu constater qu'une approche à base de règles était appropriée pour identifier les questions dans les courriels. Cette approche

a l'avantage de se généraliser à plusieurs domaines et elle ne nécessite donc pas d'adaptation compliquée. Nous avons pu constater que l'ajout de règles n'améliorait pas les résultats de l'identification, par contre, si nous utilisons les connaissances du domaine, nous croyons pouvoir en améliorer la précision et le rappel. Cette avenue n'a pas été examinée en détail parce que l'effort nécessaire pour spécialiser la méthode ne représentait pas vraiment un gain de productivité par rapport au nombre supplémentaire de courriels du corpus que nous aurions pu traiter.

Chapitre 5

Classification et formalisation de la question pour la sélection des sources de réponse

La sélection des documents candidat en question-réponse textuelle a pour but de limiter la quantité de données à analyser. La méthode généralement employée est de transformer la question en une requête qui est ensuite soumise à un système de recherche d'information. La recherche d'information retourne un ensemble restreint de documents ou de passages qui seront analysés en détail pour y récupérer une réponse à la question. Lors de cette étape, l'information non-pertinente à la requête est filtrée. L'élimination de l'information non-pertinente facilite la recherche de la réponse, puisque nous pouvons négliger la complexité de calcul des opérations à réaliser pour nous concentrer sur leur précision.

Lorsque nous considérons le problème de question-réponse textuelle, le filtrage de l'information (ou la sélection des documents candidats) peut se faire par la recherche d'information parce que nous travaillons à partir de l'hypothèse que la réponse est présente dans un document textuel. Cette hypothèse est renforcée par la grande quantité de documents qui nous permet d'exploiter la redondance de l'information et de supposer que la réponse est présente dans un document comme une réécriture de la question sous une forme déclarative. Dans notre problème, il y a peu de documents textuels contenant une réponse à une question et la réponse est rarement mentionnée

plusieurs fois, comme c'est le cas dans un corpus de textes volumineux. Par contre, nous cherchons la réponse à travers diverses sources et types d'information. Au lieu de restreindre la quantité de documents à analyser, nous restreignons les possibilités de traitement pour trouver la réponse, en fonction des caractéristiques de la question.

Nous pouvons chercher la réponse de plusieurs manières dans notre problème, en fait, chaque question semble nécessiter une procédure différente. Les questions peuvent être analysées à plusieurs niveaux et les caractéristiques des questions influencent le traitement qui sera effectué pour trouver la réponse. Nous pouvons regrouper les caractéristiques des questions dans des taxonomies auxquelles nous associons différents niveaux d'analyse linguistique, p. ex. syntaxique, sémantique, pragmatique. Ces taxonomies seront utilisées pour classifier les questions selon certaines caractéristiques pour restreindre le nombre de façons de trouver la réponse.

Outre la classification des questions, la représentation formelle de la question restreint aussi le nombre de façons de récupérer la réponse. La formalisation de la question est un moyen de donner une préférence à une méthodologie de réponse. Une représentation logique de la question tendra à privilégier des méthodes calculatoires strictes, tandis qu'une représentation basée sur l'acte discursif tentera de définir dans quelle perspective l'information devra être recherchée.

5.1 Classification des questions pour les services d'information

Les questions que nous considérons dans les services d'information sont différentes des questions factuelles apparaissant en question-réponse traditionnelle ; elles doivent être interprétées dans leur contexte, celui du service d'information. Le contexte du service d'information est similaire à celui des services de référence numérique. Les services de référence sont présents dans les bibliothèques et les centres de documentation, ils orientent les utilisateurs dans leur recherche d'information en leur donnant des réponses ou des moyens de trouver des réponses à leurs questions. La variante numérique du service de référence consiste à offrir le service par des moyens de communications numériques. Nous lions les services de référence numérique et les services d'information parce que l'étape de classification des requêtes dans les services de

référence numérique [52] est réalisée pour traiter la requête avec les bonnes ressources comme nous tentons de le faire dans les services d'information.

Au chapitre précédent, nous avons présenté une méthode pour identifier automatiquement les questions contenues dans les courriels. Dans les services de référence, l'aiguillage des requêtes est réalisé manuellement et la tâche d'identification de la question est réalisée intuitivement par un préposé. La classification initiale des requêtes est essentielle, elle permet aux intervenants de déléguer les questions aux personnes ayant les compétences pour y répondre, car une seule personne ne peut traiter l'ensemble des demandes acheminées au service d'information. L'aiguillage des requêtes est essentiellement une étape de classification, c'est l'étape initiale de la compréhension de la question et elle est cruciale pour le traitement efficace des demandes.

Les façons de classifier les requêtes diffèrent d'une entreprise à l'autre, les critères de classification utilisés dépendent largement de la façon dont celles-ci gèrent l'information. La portée des taxonomies dépend de l'information que le service doit traiter. Dans les centres d'information spécialisés, les taxonomies sont restreintes au domaine de traitement du service. Lorsque le service d'information est centralisé et qu'il doit traiter plusieurs types de requêtes différentes, plusieurs taxonomies analysant différents aspects de la requête doivent être exploitées pour optimiser le traitement des requêtes. L'étude des taxonomies de questions nous permet de regrouper les schémas de classification dans des catégories correspondant au niveau d'analyse linguistique de la question et au type de traitement linguistique que nous devons réaliser pour répondre à la question.

L'objet de la question n'est pas exprimé de la même manière en question-réponse que dans les services d'information. Dans la question-réponse factuelle, les questions peuvent être traitées par une analyse partielle du texte, principalement à partir de techniques d'extraction d'information, puisque la réponse provient de textes contenant des suites de faits, comme les articles de presse, où la structure informative suit un modèle traditionnel. Dans les services d'information, la réponse provient de sources diverses, d'où la nécessité de sélectionner des sources d'information plus efficacement. Dans les deux cas, la classification de la question est une étape de traitement facilitant l'extraction de la réponse tout en restreignant le contexte d'analyse des documents.

Dans les services d'information, la question constitue l'élément initiateur d'un dialogue avec un répondant. Dans un courriel les possibilités de rétroaction de la part du demandeur sont faibles. L'identification de l'information contextuelle de la question est une tâche complexe, ainsi, en évaluant la requête sous différents aspects nous pouvons identifier les traitements qui sont nécessaires pour répondre au courriel.

La classification des questions nous sert à identifier les meilleurs outils pour traiter correctement la question. Les tâches cognitives que nous réalisons sans cesse sont structurées et nécessitent des traitements différents en fonction du langage, du raisonnement et du modèle des problèmes que nous résolvons. De manière analogue aux systèmes mécaniques et électriques qui utilisent différents modèles pour représenter le même monde physique. La classification des questions permet d'identifier le type des questions et de leur associer le modèle utilisé pour y répondre. En associant une question à la bonne ressource, nous augmentons les chances d'y trouver une bonne réponse.

La classification des questions dans les services d'information où le traitement est réalisé par des personnes, se divise en trois étapes :

1. séparer les questions légitimes des autres communications ;
2. séparer les questions pertinentes des questions ne faisant pas partie du domaine de référence ;
3. déterminer la ressource la plus appropriée pour répondre à chacune des questions.

Nous avons déjà abordé, dans le chapitre précédent, les deux premières étapes de traitement sous la perspective de l'identification de la question. Nous n'avons pas été jusqu'à identifier automatiquement la pertinence de la question, mais nous avons dû le faire manuellement lors de l'analyse du corpus pour déterminer les questions que nous considérons dans le problème. Dans notre cas, déterminer si un segment de texte est une question ou non, est une forme de classification à partir de données syntaxiques. Les questions se classifient selon plusieurs critères liés au niveau d'analyse linguistique que nous utilisons.

Un autre exemple de traitement linguistique simple est l'identification des pourriels. Généralement, seulement en jetant un coup d'oeil au courriel, une personne peut déterminer à partir des caractéristiques de surface du courriel s'il constitue un pourriel ou non. Ce traitement linguistique peu évolué est du même niveau que les

algorithmes d'identification des pourriels qui utilisent des méthodes de statistique bayésienne sur la fréquence des mots.

Notre analyse d'un texte est influencée par la tâche que nous réalisons, car le degré d'analyse est lié à la difficulté de la tâche d'interprétation de l'information. L'analyse d'une question par la classification est un processus cognitif qui est parfois réalisé inconsciemment pour nous permettre de comprendre la question. Les personne utilisant le courriel font souvent cette classification pour passer d'un sujet à l'autre et pour changer de contexte d'interprétation du courriel.

Puisque la classification des questions est déjà réalisée dans les services d'information et qu'elle nous apparaît essentielle pour analyser les questions, nous identifions des taxonomies de questions correspondant à des niveaux d'analyse linguistique que nous pouvons exploiter pour solutionner notre problème. Dans le tableau 5.1, nous retrouvons les taxonomies de questions qui influencent la façon dont nous traitons les questions dans les courriels et que nous présentons dans les sous-sections suivantes. Les taxonomies ne sont pas liées seulement au niveau d'analyse linguistique, il y a d'autres schémas de classification pour traiter l'information qui n'ont pas de fondement linguistique, mais qui dépendent du domaine d'application.

Taxonomie de questions		**Niveau d'analyse linguistique**
Wh-words	(5.1.1)	Syntaxique
Sujet de la question	(5.1.2)	Sémantique
Fonction de la réponse attendue	(5.1.3)	Analyse du discours
Forme de la réponse attendue	(5.1.4)	Pragmatique
Types de sources dont la réponse peut être extraite	(5.1.5)	

TAB. 5.1 – Correspondance entre les taxonomies de questions et les niveaux d'analyse linguistique.

La classification des questions est une façon d'analyser la question pour en comprendre le besoin d'information et pour y répondre. Les taxonomies de questions sont essentielles au succès de la question-réponse [11, 43]. Chaque taxonomie a une utilité dans la compréhension de la question et analyse un aspect différent de la question. Les cinq taxonomies présentées constituent les facettes d'une méthode d'analyse des questions.

5.1.1 Taxonomie des *wh-words*

La classification la plus simple consiste à analyser la question selon les *wh-words* (*who, where, when, what, which, how*) qui sont habituellement utilisés pour introduire une question. En question-réponse factuelle, les questions sont habituellement formulées par des *wh-words*.

- Where would I locate Nortel exchange formula ?
- What is the phone number of ComputerShare Investor ?
- How to report tax for the average cost base on both shares ?

Cette méthode de classification des questions se réalise au niveau de l'analyse syntaxique simple. Les taxonomies basées sur les *wh-words* sont très répandues dans les systèmes de question-réponse [29, 39, 45]. Elles sont habituellement enrichies par une subdivision hiérarchique des classes en sous-classes, correspondant à une spécialisation de la question. La spécialisation de la catégorisation utilise des critères de surface de la question comme le type des entités nommées.

L'utilisation des *wh-words* pour classifier les questions dans les systèmes de question-réponse sur une base documentaire composée de textes de presse est naturelle parce que ces textes sont structurés pour rapporter l'information. L'information importante concernant l'événement rapporté est souvent dans les premiers paragraphes et l'information est structurée autour : de l'événement (quoi), des personnes (qui), du moment (quand), du lieu (où), de la manière (comment) et de la raison (pourquoi). Cette façon de rédiger est efficace et elle est valide autant en français qu'en anglais.

Cette taxonomie simple correspond à un principe clair de récolte d'information. Il y a par contre deux inconvénients à utiliser cette taxonomie pour classifier les questions dans le cadre des services d'information, les questions ne débutent pas toujours par un *wh-word*, mais les *wh-words* sont aussi utilisés pour débuter des phrases qui ne sont pas des questions. Par exemple, la phrase

We were hoping that you could give us the market share in which Bell holds.

est énoncée comme une proposition, mais elle doit être comprise comme une question. (Le mot*which* n'est pas utilisé pour énoncer une question dans cet exemple.) Tandis que la phrase

What can you say ?

a la syntaxe d'une question, mais elle est généralement utilisée pour exprimer une émotion, comme un état de surprise. Ce type de taxonomie étant utilisé dans les systèmes de question-réponse, la plupart d'entre eux combinent d'autres types de classification sur le sujet de la question ou sur le type de réponse attendu, pour combler cet inconvénient.

5.1.2 Taxonomie selon le sujet de la question

La classification des questions selon leur sujet n'est pas une idée nouvelle, les bibliothèques font cette classification depuis longtemps. La classification par sujet est réalisée dans plusieurs domaines pour classifier différents objets à partir du sujet principal. La classification des questions selon les sujets émane du principe que la classification du matériel informatif suit le même patron cognitif que celui des questions : il y a une relation directe entre le sujet de la question et celui des documents.

Les taxonomies selon le sujet de la question sont adaptées à chaque service d'information en fonction de l'organisation du service et des méthodologies de classement de l'information. Dans les systèmes de question-réponse, les sujets de la question sont identifiés à partir des structures lexicales et syntaxiques de la question. Les classes de questions identifiées dans ces taxonomies sont similaires à celles retrouvées dans les subdivisions des taxonomies orientées sur les *wh-words* dans la question-réponse.

Ce type de taxonomie permet de limiter le domaine recherché à un sous-ensemble du domaine de référence. Ce résultat est atteignable en autant que le sujet de la question est classifiable de la même façon que les documents de référence. Dans les services de référence, cette taxonomie est régulièrement utilisée, son efficacité dépend de la connaissance, par les personnes offrant les services, des références disponibles. Cette taxonomie est assez souple pour accommoder la plupart des types de questions spécifiques à un service.

5.1.3 Taxonomie selon la fonction de la réponse attendue

L'analyse du type de réponse à formuler suite à l'analyse d'une question nous donne une meilleure idée de l'information devant être récupérée pour répondre à la question. La fonction de la réponse attendue spécifie le besoin d'information comblé

par la réponse, cette fonction est similaire aux types sémantiques utilisés dans les systèmes de question-réponse factuels. Par exemple dans la question *Quelle est la distance d'un marathon ?*, la fonction de la réponse attendue est la spécification d'un attribut et le type sémantique de la réponse est une distance.

Les travaux réalisés par Wendy Lenhert [41] avec le système QUALM pour la compréhension d'histoire sont les précurseurs de cette taxonomie. La taxonomie proposée par Lenhert a été développée pour le problème de la compréhension d'histoire et a ensuite été modifiée par Graesser pour analyser des questions générales. Cette taxonomie est divisée en deux catégories selon la longueur de la réponse attendue, courte ou longue (tab. 5.2)

Cette taxonomie est en fait une classification mixte de la question et de la réponse. Dans les systèmes de question-réponse, les paires formées de la question et de la réponse sont utilisées pour déterminer les classes de questions en fonction des caractéristiques sémantiques des réponses. Cette classification des questions dans les systèmes de question-réponse est différente de celle présentée ci-dessus, puisqu'elle est effectuée à partir de la réponse et en fonction des données qui peuvent être extraites dans les documents de référence. La taxonomie selon la fonction de la réponse attendue ne se définit pas uniquement en fonction de la syntaxe ou de la sémantique de la question. Pour utiliser cette taxonomie dans les services d'information, nous devons analyser la question dans le contexte du discours.

5.1.4 Taxonomie sur la forme de la réponse attendue

La taxonomie sur la forme de la réponse attendue s'oriente autour de la façon dont la réponse est effectuée, dans les bibliothèques cette taxonomie est utilisée pour classifier les transactions. Une manière élémentaire de classifier les transactions consiste à séparer les transactions de référence des transactions de direction. Une transaction de référence nécessite un traitement de l'information pour donner une réponse au requérant tandis qu'une transaction de direction aide une personne à trouver l'information qu'elle recherche à partir des ressources à sa disposition. Ces deux classes de réponse sont très vastes, c'est pourquoi elles ont été subdivisées, formant ainsi une taxonomie, pour une meilleure compréhension des procédés menant à une réponse. Plusieurs taxonomies sur ce thème ont été développées selon l'approche des services.

Taxonomie des réponses courtes

Vérification	Est-ce qu'un fait est vrai ? Est-ce qu'un événement s'est produit ?
Choix multiples	Est-ce X ou Y ? Lequel de X, Y ou Z ?
Complétion de concepts	Qui ? Quoi ? Quand ? Où ? Quelle est la référence au nom ?
Spécification d'attribut	Quel qualificatif est assigné à l'entité X ?
Quantification	Quelle est la valeur d'une variable quantifiable ? Combien ?

Taxonomie des réponses longues

Définition	Qu'est-ce que X ?
Exemple	Qu'est-ce qui est une instance ou une valeur possible d'une catégorie ?
Comparaison	Qu'est-ce qui distingue ou « unit » les items X et Y.
Interprétation	Qu'est-ce que l'on peut déduire d'un ensemble de données ?
Cause	Quel état ou événement est à la source d'un événement ou d'un état ?
Effet	Quelle est la conséquence d'un événement ou d'un état ?
But	Quels sont les motifs et les buts menant à la réalisation d'une action ?
Démarche	Comment l'agent a pu accomplir son but ? Quel était son plan ?
« Enablement »	Qu'est-ce qui a permis à l'agent de réaliser une action ?
« Expectational »	Pourquoi un événement attendu ne s'est pas produit ?
Jugement	Quelle valeur le répondant donne à une idée ou un conseil ?
Assertion	L'émetteur énonce son manque de connaissance ou son incompréhension d'un concept.
Directive	L'émetteur veut que le répondant réalise une action.

TAB. 5.2 – Taxonomie des réponses courtes et longues

Ce type de classification est nécessaire à la compréhension de la question parce que cette classification influence le type de traitement réalisé pour répondre à la question.

Les classes de la taxonomie que nous présentons permettent beaucoup de souplesse pour la classification des questions. À ce niveau la classification est dépendante de la façon dont l'information est traitée par l'organisation. Nous pouvons classifier les formes de réponses dans les catégories présentées dans le tableau 5.3. Pour chaque catégorie, nous y décrivons la forme de la réponse attendue et un exemple de question tiré du corpus BCE-4.

La classification de la question selon cette taxonomie n'est généralement pas réalisée consciemment par le préposé chargé de traiter la question, par contre, elle est essentielle, car elle nous informe sur le type de traitement devant être effectué pour générer une réponse satisfaisante. Lors d'une requête, le préposé doit être en mesure d'analyser l'information demandée. Lorsqu'il a compris ce qu'il doit retourner, il peut alors élaborer une stratégie pour répondre à la requête. L'analyse de la question sur la forme de la réponse donne au préposé un aperçu du travail qu'il doit faire pour la recherche et la rédaction de la réponse.

5.1.5 Taxonomie selon le type de sources

Le traitement des questions doit aussi considérer les types de sources d'information à consulter pour extraire la réponse. La taxonomie des questions selon le type de sources s'appuie sur l'idée énoncée dans la taxonomie précédente qu'une partie du processus de compréhension de la question consiste à trouver les sources d'informations nécessaires à la rédaction d'une réponse.

Par exemple, si nous traitons une question factuelle concernant la géographie, alors la ressource appropriée est un atlas. Lorsque nous répondons à une question à propos des cotes boursières, nous devons consulter un registre de données boursières. Quand la réponse nécessite un traitement plus complexe, nous devons utiliser plusieurs types de sources pour répondre à la question.

L'émetteur de la question a parfois une idée du format de la réponse et d'où elle peut être extraite, dans ce cas, la tâche du préposé est de le guider vers les sources pertinentes. Lorsque les préposés ne connaissent pas parfaitement l'information contenue dans les sources à leur disposition, il faut pouvoir acheminer la requête à des personnes expérimentées avec les sources pertinentes.

Direction	Localisation d'une source d'information spécifique.
	Where can I find information on the terms of the new Series T Preferred shares which I believe will have a fixed rate of dividend ?
Possession	Demander si l'information est disponible ou si l'organisation possède une source d'information permettant de répondre à la question.
	Do you have the Valuation Day (1971 ?) share price of BCE ?
Référence prête-à-utiliser	Question simple avec une réponse factuelle accessible directement d'une source d'information.
	I would like to know the price of Bell Canada Stock for December 31, 1969.
Reproduction	La reproduction exacte d'une source d'information originale.
	Please send me how to calculate the Average cost base for BCE and NTL for tax purpose.
Description	La définition d'une entité, généralement plus brève que la source d'information (résumé).
	Please give me the definition of the Calculated Trading Price for the purpose of determining the dividend payments of BCE Preferred Series S.
Conseil	Conseil sur le choix de la source de l'information.
	I was just wondering who I could contact to update the website or who to discuss the issue with.
Instruction	Utilisation des sources d'informations disponibles.
	Could you please tell me how to gain an access to the highest and lowest stock prices of BCE on a daily basis during the period of March, April, and May, 2000.
Recherche	Nécessite un certain effort et l'utilisation de plusieurs sources pour être en mesure de synthétiser une réponse.
	I purchased $10 000 worth of shares in Excel, what are they now worth and what is the new stock symbol.
Citation	Liste de références sur un sujet
	I would appreciate a response giving some detailed answers regarding the planned sale of BCI and the conditions of sale that will apply.
Analyse	Analyse de données, par exemple une tendance, un argument pour ou contre, une relation de cause à effet ou une comparaison.
	Can you please comment on the rationale for Teleglobe name change, your thoughts on Moody's review for downgrade for Teleglobe and what your committment to this entity will be.
Critique	Question nécessitant une analyse subjective d'un sujet.

TAB. 5.3 – Taxonomies des questions sur la forme de la réponse attendue

La classification des questions par les types de sources est dépendante de la connaissance des collections de documents et de l'expérience du préposé. Ces connaissances sont difficiles à intégrer dans un système informatique, mais les types de sources sont bien définies dans les systèmes d'information. Ces taxonomies ne sont pas aussi bien définies que les précédentes parce qu'elles sont liées à l'organisation de l'information et à la classification des documents à l'intérieur de l'entreprise.

5.2 Formalisation de la question

Dans la section précédente, nous avons analysé les questions de manière pragmatique, en les classifiant selon différents aspects. Dans cette section, nous considérons les caractéristiques de la question dans un cadre formel. Les taxonomies de questions correspondent toutes à un certain niveau d'analyse formelle de la question.

Une question se caractérise par ses propriétés syntaxiques et grammaticales (l'ordre des mots, l'occurrence d'un pronom interrogatif et la présence du point d'interrogation) et par son rôle dans le discours en tant qu'acte communicatif, en énonçant une requête contenant une information indiquant l'attente d'une réponse. Un acte interrogatif peut aussi avoir comme conséquence la réalisation d'une action. Les questions sont généralement énoncées dans le but d'avoir une réponse, mais il y a des cas où la forme interrogative est utilisée pour initier une discussion ou pour valider une opinion sous la forme d'un argument rhétorique. Dans nos travaux, nous nous intéressons qu'aux questions nécessitant une réponse ou une action de la part du répondant, les autres manifestations de la question n'apparaissant pas dans le cadre des services d'information.

En linguistique, l'étude des phrases interrogatives appartient au domaine de la syntaxe tandis que l'étude de l'acte interrogatif appartient au domaine de la pragmatique. Les approches pour étudier la question se catégorisent selon l'orientation vers le développement d'un modèle logique *pur*, sans lien avec l'aspect linguistique, ou selon une approche linguistique, dans des formalismes similaires aux grammaires de Montague. Aucun modèle de la question ne semble être en mesure de faire consensus [25], les approches de la question que nous présentons sont celles que nous jugeons potentiellement utiles pour analyser la question dans le cadre de la question-réponse.

5.2.1 Logique

L'approche logique de la question tente de réduire la compréhension d'une question à une valeur de vérité. Les premiers travaux donnant un cadre théorique à la question proviennent de la logique érotétique [25, 55], dans le but de fournir un ensemble de méthodes formelles pour analyser et formaliser les concepts de question et de réponse. Cette façon de concevoir la question et la réponse se compare à ce qui est fait en recherche d'information ou même en question-réponse. Selon cette méthode d'analyse, les questions sont simplement définies comme *les phrases qui nécessitent une réponse.*

Selon cette logique, la question se décompose en deux parties : le sujet et la requête. Le sujet étant les états possibles du monde, tel que sous-entendu par la question, et non pas le sujet au sens grammatical. La requête décrit l'état du monde, tel qu'il est exigé par la réponse. Dans cette théorie, une question est vraie ou fausse si la réponse à la question est un énoncé vrai ou non. Par contre, certaines réponses ne peuvent pas être simplement vraies ou fausses, c'est le cas d'une réponse à une question faisant référence à une présupposition fausse. Ce modèle de la question est difficile à exploiter parce qu'il est trop rigide et les calculs qui y sont effectués sont complexes. De plus, cette formalisation est difficile à intégrer dans une interaction avec un utilisateur, sauf pour des questions calculatoires simples.

5.2.2 Pragmatique

Nous abordons l'aspect pragmatique de la question du point de vue de l'acte de parole (*speech act*) qui considère l'utilisation du langage dans le but d'accomplir une tâche. Selon cette approche, l'acte de parole est l'unité principale d'analyse sémantique. Chaque phrase ou chaque expression doit être analysée en fonction du rôle qui lui est affecté dans l'acte de parole. L'acte de parole est le moyen utilisé pour réaliser un acte en utilisant le langage.

Il y a quatre grandes caractérisations des actes dans la théorie de l'acte de parole, développée par Austin [3] et Searle [61].

L'acte d'énonciation (ou locutoire) consiste à émettre des sons, des mots, des phrases d'un langage. Cette action n'est pas très intéressante à étudier d'un point

de vue pragmatique parce qu'elle ne communique rien. Elle pourrait être réalisée par un perroquet ou un ruban magnétique qui répète ce qu'il a enregistré. Les *sons* émis au cours de cet acte ont une signification, mais pas nécessairement la même, pour les participants à l'acte de parole et ne produisent pas d'effet. Un acte locutoire peut être un chuchotement, un cri ou un murmure. L'intérêt de l'acte locutoire est qu'habituellement sa réalisation est aussi un acte illocutoire ou un acte perlocutoire.

L'acte illocutoire se réalise en énonçant quelque chose, cet acte doit avoir le même sens pour l'émetteur et le récepteur du message. La force de l'acte illocutoire consiste à avoir une signification sémantique comparativement à l'acte d'énonciation qui n'a qu'un sens syntaxique. Les actions réalisées lors d'un acte illocutoire peuvent être : promettre, demander, suggérer, rapporter, énoncer, etc.

L'acte perlocutoire se réalise en énonçant quelque chose qui produira un effet sur les sentiments, les pensées ou les actions de l'audience. La réalisation d'un acte perlocutoire se fait en réalisant un acte illocutoire, mais avec une intention et dans une circonstance en particulier. Dans l'acte illocutoire, la compréhension du message entre les participants est un point important, cependant lors de l'acte perlocutoire, le message doit être compris et accepté par le récepteur. L'acte illocutoire n'est pas réalisé explicitement dans le message, on ne convainc pas explicitement quelqu'un, même si c'est l'effet que l'on veut avoir par cet acte de parole.

L'acte de proposition consiste à exprimer des faits lors d'un acte de parole, en donnant un nom à une entité ou en la qualifiant. Par exemple, en énonçant la phrase « Alex est blessé. », l'énonciateur identifie l'entité Alex avec le nom « Alex » et le caractérise par le prédicat « est blessé ». Le contenu propositionnel d'un acte ne détermine pas l'aspect qu'il prend, plusieurs actes illocutoires peuvent partager le même contenu propositionnel, mais ne pas avoir la même signification.

Dans l'acte de parole, la question a une force provenant du fait qu'elle est un acte illocutoire. La personne qui pose une question attend une réponse, cette attente est le résultat de l'acte de parole. Dans le cadre des services d'information, les questions sont des actes perlocutoires, puisque l'émetteur de la question s'attend à ce que la personne qui la reçoit réagisse en lui donnant une réponse. Comparativement à l'approche logique, cette approche distingue les conditions de succès et de satisfaction. La condition de succès détermine si l'émetteur, par son acte illocutoire, a réussi à

communiquer son besoin. Cette condition est dépendante de la force que l'émetteur a pu transmettre lors de l'acte illocutoire. La condition de satisfaction dépend du contenu propositionnel de la question qui est déterminé par le fait que la réponse est vraie ou fausse. En ce qui nous concerne, la caractérisation des actes est importante pour identifier les courriels nécessitant des réponses.

L'acte de parole est une composante essentielle dans l'analyse du discours. L'analyse du discours considère les actes de parole dans le contexte d'une conversation, comparativement à l'acte de parole qui considère l'acte comme une unité auto-suffisante. En analyse du discours, le destinataire du message est actif dans la définition du message, car la rétroaction doit être acceptée par l'émetteur. Dans ce type d'analyse, les interlocuteurs doivent s'entendre sur la signification du message. De la théorie développée en analyse du discours, nous pouvons supposer qu'une partie des correspondances d'un service d'information nécessite que l'émetteur initial ait à préciser le sens de sa requête. Parfois, un acte de parole n'est pas réalisé par une phrase interrogative, malgré cela, à cause du contexte du service d'information, nous devons l'interpréter comme une question.

5.2.3 Sémantique

L'approche pragmatique considère les questions sous la forme d'un acte de parole, une autre façon d'étudier la question est d'analyser le contenu sémantique de la question. Selon le modèle sémantique de la question, une question est un objet particulier qui doit être étudié en considérant les réponses possibles. L'approche générale proposée par Hamblin consiste à considérer la question et sa réponse selon trois principes [26] :

1. la réponse à une question est une phrase ou une assertion ;
2. les réponses possibles à une question constituent un ensemble exhaustif de possibilités mutuellement exclusives ;
3. connaître le sens d'une question correspond à connaître une réponse possible à cette question.

Cette façon de considérer la question détourne notre attention de l'analyse syntaxique de la question vers une analyse de la question en fonction de la réponse. Le premier principe donne un rôle à la réponse, celui de fournir de l'information,

la sémantique de la question peut alors être évaluée comme un propos. Le second principe spécifie que les propos qui sont des réponses à une question, sont exclusifs les uns par rapport aux autres, la véracité d'une réponse implique que les autres sont nécessairement fausses. Donc si une réponse est vraie, elle définit complètement et précisément l'information demandée par la question. Ces deux principes font en sorte que l'ensemble des réponses possibles à une question constitue une partition de l'espace des réponses logiques. Le troisième principe vient clarifier la situation en statuant que la signification d'une question est en fait l'ensemble des réponses logiques possibles. Il émane des trois principes de Hamblin, une unique représentation logique de la question exprimée en fonction d'un ensemble de propositions.

5.2.4 Besoin d'information

Le besoin d'information est une façon d'aborder l'étude de la question dans les services de référence et en science de l'information. Ce point de vue se définit intuitivement par le fait qu'une personne consultant un service d'information a nécessairement un besoin d'information. Le besoin d'information est difficile à exprimer clairement parce qu'il est motivé par plusieurs facteurs. La personne qui cherche à répondre à son besoin d'information fait face à une anomalie ou à une brèche dans son état des connaissances, en fonction d'un sujet ou d'une situation. Elle n'est pas toujours en mesure d'exprimer spécifiquement l'information nécessaire pour compléter ses connaissances, ce qui explique pourquoi il est difficile de répondre correctement aux questions.

Cette façon d'étudier la question consiste à évaluer la question en fonction de l'information qu'on peut fournir comme réponse. L'information fournie sera évaluée selon qu'elle aide ou non à désambiguïser le besoin d'information présent dans la question. Une question est ainsi considérée comme une action ayant pour but de combler un besoin d'information correspondant à une brèche dans la compréhension d'un sujet concernant l'univers qui l'entoure. Le besoin d'information est en fait un phénomène observable dans un acte du discours ayant la force d'amener une personne à donner une réponse dans le cadre d'une conversation.

5.3 Conclusion

La classification et la formalisation des questions influent la récupération de la réponse. Dans le cadre de l'automatisation d'un système de réponse au courriel, la classification des courriels apparait être une étape de traitement naturelle. Dans notre cas, nous utilisons les taxonomies de questions pour mettre en évidence les traitements qui sont réalisés pour analyser une question, sans vouloir classifier explicitement chaque courriel.

Lorsque nous devons répondre à un courriel, l'identification de la question par l'analyse des caractéristiques syntaxiques des courriels est la première étape à réaliser. La taxonomie des *wh-words* est alors appropriée pour identifier les questions. Les données avec lesquelles nous travaillons ne nous permettent pas d'exploiter les *wh-words* efficacement, car plus de la moitié des questions sont écrites sans *wh-words*. La taxonomie des *wh-words* est le point de départ de la grammaire de surface pour identifier les questions dans les courriels que nous présentons au chapitre précédent.

Lors du traitement manuel des requêtes, les préposés ont l'habitude de classifier les courriels dans des catégories que nous pouvons assimiler à des sujets. Dans le traitement automatisé des courriels pour le service de relations avec les investisseurs, la classification automatique est difficile à réaliser. Suite aux travaux de Julien Dubois [20], nous avons choisi de ne pas classifier automatiquement les courriels par des méthodes quantitatives parce que les critères de séparation des classes de courriel n'apparaissent pas naturellement dans les requêtes. La distribution des mots dans les courriels n'est pas assez différente pour distinguer les classes de courriels portant sur le même sujet. Les mots fréquents le sont dans toutes les catégories et la plupart des mots ont une fréquence d'apparition peu élevée, de sorte que les mots pertinents ne peuvent être différenciés des mots ordinaires. Puisque la classification des questions ne peut être réalisée en se basant sur le modèle de la fréquence de mots, nous utilisons des caractéristiques linguistiques plus évoluées pour exploiter l'information contenue dans le courriel.

Les requêtes formulées dans les courriels ne sont pas toujours des questions avec comme réponse une entité nommée, comme c'est généralement le cas dans la question-réponse factuelle. La personne répondant à l'utilisateur doit évaluer le rôle de l'information recherchée dans la quête d'information de l'utilisateur. La taxonomie des

questions selon la fonction de la réponse sert à déterminer ce que l'utilisateur recherche. La fonction de la réponse à la question nous permet de choisir la façon dont nous cherchons l'information et rédigeons la réponse. Pour notre problème, cette approche est difficile à exploiter puisque nous devons analyser en détail la structure du message et, lorsque nous devons exploiter cette taxonomie, la spécification du contexte est généralement incomplète. L'autre réticence que nous avons à exploiter cette méthode pour classifier les questions est l'imprécision des propriétés linguistiques des questions qui servent à établir clairement le type d'une question.

Dans les systèmes de question-réponse factuels, les taxonomies de questions ne considèrent pas un niveau d'analyse linguistique plus évolué qu'une analyse sémantique partielle. Notre problème étant plus complexe, nous devons analyser la question plus en profondeur en considérant l'aspect pragmatique, puisque nous pouvons comparer cette analyse avec les traitements effectués dans les services d'information. Cette perspective nous amène à considérer deux aspects de la question qui concernent directement le mécanisme de réponse : la forme de la réponse, qui nous permet d'identifier l'information que nous recherchons et que nous retournons en réponse, et le type de source à consulter, qui nous permet de savoir où aller chercher la réponse et quels raisonnements nous devons réaliser pour y arriver.

Chapitre 6

Analyse des documents candidats

Les façons d'analyser les documents candidats en question-réponse factuelle sont nombreuses. La diversité des approches provient de la similarité entre l'analyse des documents candidats et les domaines de l'extraction d'information et de la génération de résumé. Dans toutes les approches, le but de l'analyse des documents candidats reste toujours le même, identifier l'information pertinente pour répondre à la question à partir des données sélectionnées précédemment.

Dans notre problème, les données textuelles ont une utilité restreinte parce qu'il n'y a pas assez de documents à analyser et les réponses aux questions du corpus BCE-4 ne peuvent généralement pas être extraites du corpus de référence, constitué du site web principal et des communiqués de presse de l'entreprise BCE. Par contre, lorsque nous pouvons récupérer la réponse des données textuelles, l'analyse des documents candidats devient essentielle. Puisque le système pour lequel nous présentons une architecture doit récupérer des réponses de plusieurs types de source, la méthode d'analyse des documents que nous décrivons est multifonctionnelle. Nous l'utilisons pour analyser les documents du corpus et les courriels contenant les questions.

Notre méthode d'analyse des documents consiste à identifier les actions rapportées, pour ensuite repérer les entités ayant un rôle informationnel dans la phrase, tels que : le sujet, l'objet et les compléments de l'action. Suite à cette analyse, nous pouvons constituer une représentation sous la forme de *frame* allégé. Pour repérer les rôles sémantiques, notre méthode utilise une approche statistique de classification des constituants de la phrase, que nous avons déjà présenté lors des rencontres de la

Société francophone de classification [7] (même si les données de notre travail sont en anglais). Le repérage des actions et des rôles sémantiques est utilisé pour traiter les documents candidats, autant que les questions. L'avantage d'analyser la question et les sources avec les mêmes outils est d'avoir accès à la même représentation des données, ce qui permet d'utiliser la même logique de traitement pour toutes les étapes.

Le traitement précis de l'information est nécessaire dans notre problème parce que les mêmes entités et les mêmes actions sont régulièrement mentionnées dans les textes et les questions, sans qu'elles y jouent le même rôle. En comparaison avec la question-réponse factuelle, l'information ne doit pas seulement être analysée selon le type des entités nommées. Nous ne pouvons nous contenter d'identifier un lieu ou une personne, le contexte dans lequel ils sont utilisés détermine leur sens. Par exemple, si une entreprise vend ou achète une autre entreprise, nous devons déterminer avec exactitude laquelle est acheteuse, laquelle est vendeuse et si possible l'information complémentaire à cette action. La méthode d'analyse que nous présentons s'attarde justement à traiter ce problème pour les documents candidats et les questions.

6.1 Structures prédicat-arguments

Notre mise en oeuvre de l'analyse des documents consiste à repérer les structures prédicat-arguments pour déterminer le rôle sémantique des arguments d'un verbe. Cette tâche est complexe, l'identification des événements et des acteurs en cause est un des objectifs principaux en extraction d'information. Les rôles sémantiques sont des modélisations linguistiques très courantes et importantes lorsqu'on considère l'aspect de la compréhension du langage. Les structures prédicat-arguments donnent une interprétation sémantique des phrases en déterminant qui a fait quoi, à qui, où, quand, comment et pourquoi.

La méthode choisie pour représenter les actions et les rôles est similaire à la représentation en frame, où le verbe est l'élément principal auquel les arguments complétant l'information sont associés. Par exemple, dans la perspective de FrameNet [4, 60], la phrase « She blames the Government for failing to do enough to help. » contient une instance du frame JUDGEMENT. Nous associons au frame JUDGEMENT

les rôles JUDGE, EVALUEE et REASON, pour chaque partie de la phrase, identifiées de la façon suivante :

> [$_{Judge}$ She] **blames** [$_{Evaluee}$ the Government] [$_{Reason}$ for failing to do enough to help].

Dans notre problème, nous traitons seulement les actions réalisées par des verbes. Nous pourrions étudier les actions réalisées par d'autres mots que les verbes en considérant une initiative comme TimeML [9] ou NomBank [44], mais les ressources nécessaires pour réaliser notre approche n'étaient pas exploitables. Nous utilisons plutôt le corpus Proposition Bank (PropBank) [35, 48] qui introduit les structures prédicat-arguments, une structure sémantique plus générale que les frames. Nous ne traitons pas non plus le français parce que des ressources comme le PropBank ne sont pas encore disponibles en français et que 80% des courriels dont nous disposons pour le projet Merkure de réponse automatisée aux courriels sont rédigés en anglais.

6.1.1 Proposition Bank

La représentation des rôles sémantiques par le PropBank est orientée vers une approche pratique de l'identification des rôles sémantiques. Le PropBank est un corpus d'annotation qui ajoute de l'information sur les rôles sémantiques aux structures syntaxiques du Penn TreeBank. La façon de représenter les rôles sémantiques du PropBank a l'avantage de couvrir tous les verbes du Penn TreeBank, dans le sens où ils sont utilisés dans le corpus, et la taille du corpus permet d'obtenir des données statistiques représentatives. Par contre, cette représentation ne tient pas compte des problèmes compliqués tels que la résolution des coréférences et la désambigüisation des quantificateurs.

Pour chaque verbe présent dans le Penn TreeBank, le PropBank définit un ensemble de frames, dans lequel un frame décrit un sens d'utilisation du verbe tel qu'on le retrouve dans le corpus. Dans FrameNet, les rôles sémantiques des arguments du verbe sont nommées, par exemple : *agent*, *evaluee*, *instrument*, *judge*, *reason*, etc. Pour faciliter l'exploitation des données, le PropBank étiquette les arguments de façon générique par une étiquette préfixée par Arg et suivie d'un chiffre de 0 à 4. Ainsi, les arguments sémantiques étiquetés par Arg0 ont généralement le rôle du sujet, ou de l'agent principal de l'action, et les arguments Arg1 sont les objets de l'action,

habituellement identifiés comme le patient ou le thème dans FrameNet. Nous étiquetons les compléments adverbiaux par ArgM suivi du type d'information qu'il amène à l'action. Il y a 12 étiquettes secondaires possibles pour les compléments adverbiaux.

Prenons la question *How can I purchase Bell Canada Bonds issued at 7.% ?*, provenant du corpus BCE-4, elle contient deux prédicats : *purchase* et *issue*. Les frames décrivant la sémantique des arguments pour ces deux verbes sont détaillés dans la figure 6.1, telle que décrite dans le PropBank. Puisque nous avons deux

purchase

Arg0 :	purchaser
Arg1 :	thing purchased
Arg2 :	seller
Arg3 :	prices paid
Arg4 :	benefactive

issue

Arg0 :	issued
Arg1 :	thing issued
Arg2 :	issued to
Arg3 :	attribute, issued as or at
Arg4 :	–

FIG. 6.1 – Frames des predicats *purchase* et *issue*

prédicats dans la question, nous avons aussi deux données d'annotation, la première en fonction du prédicat *purchase* et la seconde en fonction du prédicat *issue* :

1. How can [$_{Arg0}$ I] *purchase* [$_{Arg1}$ Bell Canada Bonds issued at 7.0%] ?
2. How can I purchase [$_{Arg0}$ Bell Canada] [$_{Arg1}$ Bonds] *issued* [$_{Arg3-at}$ at 7.0%] ?

Ces deux annotations combinées à l'information provenant de l'arbre de dérivation syntaxique nous donne une représentation graphique (fig. 6.2) plus informative.

Dans cet exemple, l'information provenant de l'annotation est limitée à la localisation du prédicat et des rôles dans l'arbre de dérivation syntaxique. Par contre, les données du PropBank contiennent plus d'informations que cette information de surface. Techniquement, l'annotation d'un prédicat et de ses rôles est réalisée par une seule ligne de texte dans le PropBank. Par exemple :

```
wsj/00/wsj_0020.mrg 19 8 gold issue.01 i---a 5:1-ARG0 8:0-rel
 9:2-ARG1 17:1-ARGM-TMP
```

décrit une annotation de *purchase* dans le PropBank. L'information contenue dans cette annotation se divise en deux parties : une pour le prédicat et l'autre pour les arguments. La partie suivante : `wsj/00/wsj_0020.mrg 19 8 gold issue.01 i---a`, définit le prédicat, en débutant par le chemin d'accès du fichier où nous retrouvons le

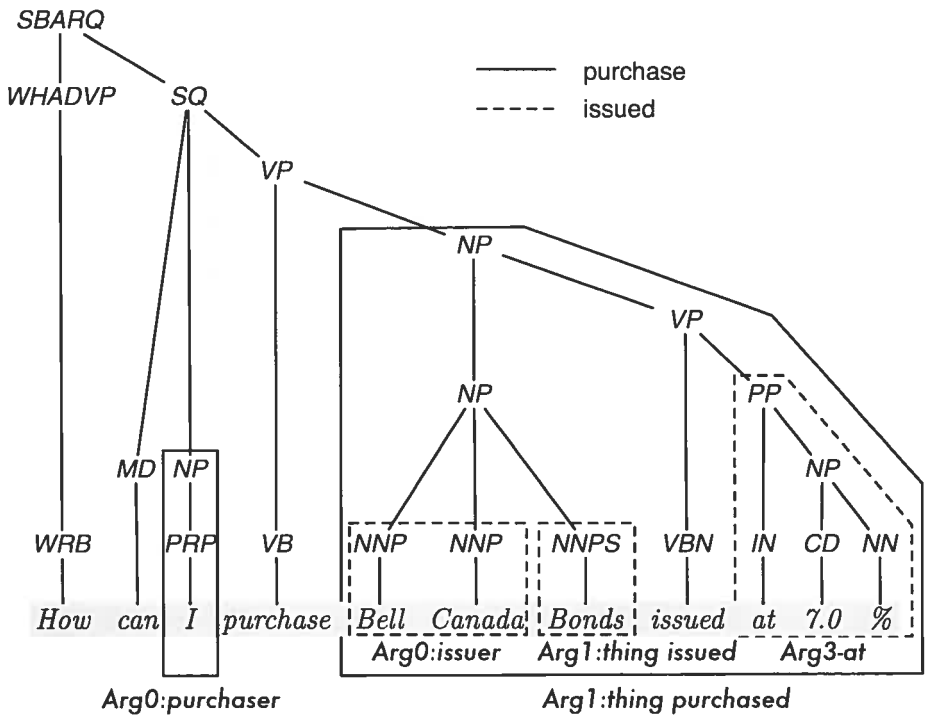

FIG. 6.2 – Analyse de la phrase *How can I purchase Bell Canada bonds issued at 7.0% ?*

prédicat (`wsj/00/wsj_0020.mrg`), suivi de la position du prédicat dans le document, 9^e mot de la 20^e phrase, les indices débutant à 0. Ensuite, nous retrouvons l'étiquette `gold` spécifiant que l'annotation a été vérifiée, le prédicat annoté `issue.01` (à l'infinitif et suffixé de l'indice de l'ensemble de rôle utilisé) suivi de l'inflection du verbe `i---a`. Il y a 5 attributs qui sont considérés, correspondant aux 5 caractères (lettre ou tiret) et les valeurs qu'ils peuvent prendre sont détaillées dans le tableau 6.1. La

form	**tense**	**aspect**	**person**	**voice**
i = infinitive	f = future	p = perfect	3 = 3rd person	a = active
g = gerund	p = past	o = progressive		p = passive
p = participle	n = present	b = both perfect		
v = finite		and progressive		

TAB. 6.1 – Valeurs possibles des attributs de l'inflection d'un verbe dans le PropBank

deuxième partie de l'annotation sert à identifier les rôles dans l'arbre de dérivation

syntaxique. Dans cet exemple nous avons trois rôles (Arg0, Arg1 et ArgM-tmp), `rel` étant le prédicat. Dans l'étiquette d'un rôle (p. ex. `5:1-ARG0`) le couple numérique sert à déterminer quel noeud joue un rôle . Dans l'exemple, `5:1` pointe vers le parent immédiat (on remonte de 1 niveau) du noeud terminal d'indice 5, tandis que le rôle `9:2-ARG1` de l'exemple pointe vers le parent du parent (on remonte de 2 niveaux) du noeud terminal d'indice 9. Le noeud terminal est toujours le plus à gauche dans l'arbre de dérivation syntaxique.

Le PropBank est le corpus que nous utilisons pour développer une méthode de repérage des rôles sémantiques pour l'analyse des documents candidats et des questions. Dans la suite du chapitre nous expliquons l'utilisation que nous faisons du corpus et comment nous solutionnons le problème de repérage des rôles sémantiques. L'avantage du PropBank par rapport à d'autres ressources est sa portée qui couvre 3 323 verbes répartis sur 112 917 annotations. De ces 3 323 verbes, 726 peuvent avoir plus d'un sens, c'est-à-dire que l'ensemble de frames du prédicat contient plus d'un frame, pour un total de 4659 sens de prédicat (frame) dans le corpus.

6.2 Repérage des rôles sémantiques

Nous réalisons la tâche d'analyse des documents candidats par le repérage des rôles sémantiques dans les documents. La définition de la tâche du repérage des rôles sémantiques est le résultat d'expériences antérieures qui nous ont permis d'identifier correctement ce que nous voulons obtenir par l'analyse des documents candidats, en fonction du domaine et de notre problème. Cette étape doit faciliter l'extraction de la réponse, c'est pourquoi nous tentons de lier les entités de la question à de l'information pertinente provenant des documents candidats.

Le problème traité est issu de l'étape d'identification des questions et de la sélection des sources de réponse. Dans le cas de l'identification des questions, la grammaire n'est pas assez développée pour analyser les questions de façon à découvrir l'information pertinente dans les documents candidats. Pour ce qui est de la sélection des sources de réponse, le problème est différent parce que nous ne présentons pas de méthodes concrètes pour identifier la source, mais plutôt un cadre général d'analyse. Ce cadre d'analyse dépend des types de sources à notre disposition et du type d'information que nous pouvons y retrouver.

Par le repérage des rôles sémantiques nous solutionnons des problèmes que nous ne traitons pas lors des deux étapes précédentes. L'analyse des questions et des données textuelles avec le même formalisme, celui des structures prédicat-arguments, nous donne la possibilité d'apparier les représentations sémantiques.

6.2.1 Repérage à partir de règles

Notre première tentative de repérage des rôles sémantiques, en dehors du formalisme des structures prédicat-arguments, est d'identifier les relations dans les questions pour déterminer de façon exacte l'information recherchée dans un courriel. L'approche que nous utilisons est inspirée de REES [2], un extracteur de relations et d'événements. Nous avons tout d'abord analysé un sous-ensemble de questions, pour déterminer les relations que nous voulons extraire. De cette analyse, nous avons créé un ensemble de règles sur les caractéristiques de surface du courriel pour extraire les relations. Pour avoir des résultats utilisables avec cette approche, nous avons rencontré un problème important puisque nous devons écrire autant de règles que nous avons de relations à extraire. Cette première expérience n'a pas été concluante parce que les questions des courriels ne partagent pas assez de caractéristiques communes.

Notre seconde tentative de repérer les rôles sémantiques est similaire à la précédente, mais nous utilisons des caractéristiques autres que les caractéristiques de surface. De cette façon, nous extrayons des caractéristiques partagées par plus de questions. Nous commençons donc par calculer la dérivation syntaxique d'une question, et ensuite, de la même façon que dans l'expérience précédente, nous définissons des règles d'identifications des relations entre les prédicats et les arguments. Ces règles sont définies dans le but de capturer l'information transmise par les relations entre les composantes de la phrase (les groupes nominaux, les groupes verbaux et les compléments) et les variations syntaxiques dans ces groupes. Les règles pour repérer les relations sont définies sur l'arbre de dérivation syntaxique de la question, d'une façon similaire à celle de Palmer, Rosenzweig et Cotton [49]. Ces règles sont des patrons correspondants à des structures arborescentes pouvant être appariées aux arbres de dérivation syntaxique des questions. Cette méthode est plus efficace que la précédente parce que les questions partagent plus de similitudes syntaxiques que lexicales, mais la grande diversité de formulations de questions ne permet pas d'avoir un nombre restreint de règles pour analyser plusieurs questions.

Ces deux expériences ne peuvent être utilisées dans notre problème parce que nous voulons une approche générique, ce qui n'est pas le cas avec les règles développées, et les patrons définis dans les expériences ne sont pas adaptés aux documents candidats. L'utilisation d'une méthode de repérage des rôles sémantiques à base de règles nécessiterait le développement de deux ensembles de règles lorsque nous changeons de domaine d'application : une pour les requêtes et l'autre pour les documents.

6.2.2 Repérage statistique

Les méthodes d'analyse sémantique (extraction de relations et repérage des rôles sémantiques) à base de règles sont laborieuses à réaliser, ont une portée limitée et doivent être reconfigurées lorsque nous voulons les appliquer à un domaine différent. Les méthodes statistiques de traitement de la langue naturelle font partie des techniques utilisées pour outrepasser les limites inhérentes aux approches à base de règles. Puisque nous disposons de ressources linguistiques importantes (FrameNet et PropBank), dont la taille substantielle permet d'extraire des données statiques significatives, nous proposons une méthode statistique pour solutionner notre problème.

Le repérage des rôles sémantiques selon un point de vue statistique est l'approche la plus adaptée à notre problème. Tout comme dans les mécanismes à base de règles, nous pouvons réaliser le repérage d'une multitude de façon. Ainsi, dans le reste de cette section, nous présentons les méthodes pertinentes à notre problème, pour ensuite décrire comment nous réalisons le repérage des rôles sémantiques dans la prochaine section.

Gildea et Jurafsky [21, 22] sont les précurseurs de l'utilisation des statistiques pour étiqueter les rôles sémantiques. Leur algorithme utilise la base de donnée FrameNet et environ 50 000 phrases, provenant du British National Corpus, annotées manuellement avec les frames. L'algorithme débute par l'extraction d'attributs lexicaux et syntaxiques pour chaque constituant de l'analyse syntaxique. Les analyses syntaxiques sont générées automatiquement par l'analyseur de Collins. Les attributs extraits pour chaque constituant sont :

- le type de la phrase (t) ;
- la catégorie gouvernante (gov) ;

- le chemin du noeud, représentant le constituant dans l'arbre de dérivation syntaxique, au verbe (*pt*) ;
- la position du constituant relativement au verbe (avant ou après) (*position*) ;
- la voix du verbe (active ou passive) (*voice*) ;
- le mot clé du constituant (*head word*) (*h*).

Nous décrivons plus en détail ces attributs dans la prochaine section puisque nous les utilisons aussi dans notre méthode. Ces attributs sont ensuite utilisés pour solutionner le problème en deux étapes par une technique de classification statistique. Les deux étapes sont : la découverte des constituants de la phrase qui ont un rôle sémantique et l'assignation des bons rôles sémantiques à ces constituants. Ces deux tâches sont en fait des tâches de classification des constituants de la phrase.

La technique pour solutionner ce problème de classification consiste à considérer la probabilité conditionnelle qu'un noeud de l'arbre de dérivation syntaxique ait un rôle r conditionnellement à la présence des attributs extraits précédemment.

$$P(r|t, gov, pt, position, voice, h) = \frac{\| < r, t, gov, pt, position, voice, h > \|}{\| < t, gov, pt, position, voice, h > \|}$$

Cette probabilité se calcule directement à partir du corpus, mais elle ne constitue pas un bon estimateur parce que les combinaisons d'attributs ont un petit nombre d'apparitions ou n'apparaissent pas dans le corpus. La probabilité ne peut pas non plus être calculée en considérant les attributs indépendamment, car il y a une interaction entre les ensembles d'attributs. La solution retenue pour avoir un estimateur efficace de la probabilité qu'un constituant ait un rôle sémantique, est de combiner un ensemble de probabilités conditionnelles sur un sous-ensemble d'attributs. La meilleure façon de combiner ces probabilités est de les disposer en treillis pour privilégier certaines dépendances entre les attributs en conservant des distributions conditionnelles lisses.

Puisque la tâche se divise en deux étapes, le repérage des rôles sémantiques nécessite deux estimateurs, un pour calculer la probabilité qu'un constituant a un rôle et l'autre pour calculer la probabilité que le constituant qui a un rôle ait le rôle r. Les deux estimateurs n'utilisent pas les mêmes distributions conditionnelles de probabilité, ni la même structure de treillis, parce que : les deux problèmes sont différents et les attributs n'influencent pas la tâche de la même façon dans les deux cas.

Cette approche donne de bons résultats dans la mesure où la dérivation syntaxique d'une phrase est correcte, ce qui n'est pas toujours le cas avec un analyseur automatique. La qualité de la dérivation syntaxique influence directement celle du repérage des rôles sémantiques [23, 56]. En utilisant un corpus annoté manuellement (Penn TreeBank) et un ensemble de rôles sémantiques normalisés (PropBank) le problème de la qualité de la dérivation syntaxique n'est plus un facteur. Puisque nous avons une plus grande quantité de données, les données statistiques sont mieux distribuées, donc plus représentatives.

Le PropBank est la ressource utilisée dans plusieurs travaux de repérage des rôles sémantiques [12, 15, 46, 54, 63, 72]. Les façons d'aborder le problème dans ces travaux s'inspirent toutes de l'approche utilisée par Gildea et Jurafsky [22], qui consiste principalement à séparer la tâche comme deux problèmes de classification sur des attributs linguistiques :

1. déterminer les constituants de l'analyse syntaxique qui ont un rôle sémantique ;
2. assigner un rôle sémantique aux constituants qui ont été identifiés lors de l'étape précédente.

La distinction réside dans le choix des attributs linguistiques et dans les méthodes de classification utilisées. Les attributs linguistiques énumérés précédemment (p. 91) sont toujours utilisés, mais d'autres attributs sont aussi proposés dans certaines approches [15, 53, 54, 63, 72]. La principale différence dans les méthodes est la technique de classification utilisée. Dans toutes ces approches, la classification est traitée comme un problème d'apprentissage supervisé, ce qui explique la diversité des travaux rapportés pour traiter ce problème.

La méthode de Surdeanu, Harabagiu, Williams et Aarseth [63] combine l'ensemble des attributs pour entraîner un classificateur sous la forme d'un arbre de décision C5. Chen et Rambow [15] utilisent aussi une méthode de classification basée sur les arbres de décision, mais avec l'algorithme C4.5. Le système de Chen et Ranbow est différent parce que les attributs ne sont pas extraits directement du Penn TreeBank, mais à partir d'une grammaire d'arbres adjoints (TAG), qui est, elle, extraite du Penn TreeBank. Les méthodes de Pradhan et coll. [53, 54] et de Moschitti et Bejan [46] classifient les constituants à partir des machines à vecteur de support (SVM). La distinction entre ces deux méthodes, qui utilisent le même algorithme de base, est la fonction de noyau de l'algorithme de classification. Dans le cas de Pradhan et coll.,

la fonction de noyau est un fonction polynomiale de second degré, régulièrement utilisée dans le domaine. Dans cas Moschitti et Bejan, la fonction de noyau est plus sophistiquée, c'est un noyau sémantique qui exploite la structure des données.

6.3 Réalisation du repérage des rôles sémantiques

La méthode de repérage des rôles sémantiques, que nous réalisons pour solutionner notre problème d'analyse de documents, est inspirée des approches de Gildea et Jurafsky [22] et de Pradhan et coll. [53]. La tâche essentielle, dans cette méthode, n'est pas le repérage en lui-même, mais plutôt la mise au point des classificateurs qui sont utilisés pour faire le repérage. Les grandes lignes de l'algorithme de repérage sont les mêmes pour toutes les méthodes, sauf pour l'identification des attributs et la technique de classification utilisée. Lorsque les classificateurs sont entraînés, les étapes pour repérer les rôles sémantiques sont les suivantes :

1. choisir une phrase à analyser ;
2. calculer la dérivation syntaxique de la phrase ;
3. identifier les prédicats $p \in \mathcal{P}$, où $\mathcal{P}$ est l'ensemble des prédicats ;
4. extraire les attributs pour chaque couple prédicat et constituant (noeud) de la dérivation syntaxique (on note ce couple $\langle p, a \rangle \in \mathcal{P} \times \mathcal{A}$ où a est une configuration d'éléments de l'ensemble d'attributs $\mathcal{A}$) ;
5. séparer les constituants qui ont un rôle sémantique avec un prédicat de celles qui n'en ont pas ;
6. pour chaque prédicat, assigner un rôle sémantique aux constituants identifiés lors de l'étape précédente.

6.3.1 Attributs linguistiques

Les attributs linguistiques utilisés dans nos expériences sont : (1) le type de la phrase, (2) la catégorie gouvernante, (3) le chemin dans la dérivation syntaxique du noeud au prédicat, (4) la position du noeud relativement au verbe, (5) la voix du verbe, (6) le mot clé du constituant, (7) la position du mot clé, (8) le prédicat. Ces attributs sont un bon point de départ pour réaliser le système, les travaux que nous

rapportons dans la section précédente utilisent aussi ces attributs comme point de départ. Dans notre implémentation, nous utilisons le *Stanford Parser* pour réaliser l'analyse syntaxique d'une phrase et *PB Tool* de Scott Cotton et Benjamin Snyder pour identifier les attributs. Nous présentons une définition sommaire des attributs ci-dessous.

Type de la phrase – Le type de la phrase est le type syntaxique du constituant, l'étiquette du noeud dans la dérivation syntaxique, p. ex. groupe nominal (NP), complément (PP), phrase adverbiale (ADVP), phrase (S).

Catégorie Gouvernante – Lorsque le noeud est un groupe nominal (NP), la catégorie gouvernante indique si le groupe nominal est dominé par une phrase (S) ou un groupe verbal (VP). Cet attribut est motivé par la corrélation entre le rôle sémantique et la réalisation syntaxique du sujet ou du complément d'objet direct.

Chemin du sous-arbre – Le chemin du noeud du constituant au verbe dans la dérivation syntaxique capture la relation syntaxique entre le constituant et le prédicat. Par exemple, le chemin du couple $\langle purchase, \mathsf{PP} \rangle$ de la figure 6.2 (*at 7.0 %*) est $\mathsf{PP} \uparrow \mathsf{VP} \uparrow \mathsf{NP} \uparrow \mathsf{VP} \downarrow \mathsf{VB}$, où $\uparrow$ signifie que le chemin monte dans l'arbre et $\downarrow$ que le chemin descend.

Position – Indique si le constituant est avant ou après le prédicat. Cet attribut est généralement corrélé avec la fonction grammaticale puisqu'un sujet apparaît habituellement avant le verbe et l'objet après.

Voix – La voix d'un verbe distingue s'il est actif ou passif. Cet attribut est lié au rôle sémantique parce que les compléments d'objets directs des verbes actifs sont les sujets des verbes passifs.

Mot clé du constituant – Le mot clé du constituant est un attribut lexical qui donne de l'information sur le rôle sémantique du constituant. Le mot clé d'un noeud est calculé par un ensemble de règles de façon déterministe [16].

Sous-catégorisation – Cet attribut est la règle de grammaire qui produit le prédicat. Dans l'exemple de la figure 6.2, la sous-catégorisation du prédicat *purchase* est $\mathsf{VP} \rightarrow \mathsf{VB} - \mathsf{NP}$.

6.3.2 Entraînement des classificateurs

Pour solutionner notre problème, nous avons entraîné plusieurs classificateurs utilisant l'algorithme d'apprentissage supervisée SVM. La première étape nécessite un seul classificateur pour classifier les constituants, selon qu'ils ont ou non un rôle sémantique. Cette étape est essentielle car environ 90% des noeuds d'une dérivation syntaxique n'ont pas de rôle sémantique. De plus, nous voulons avoir une précision et un rappel élevé parce que ce classificateur influence grandement les performances globales du repérage des rôles sémantiques. Notre deuxième tâche de classification est d'assigner le bon rôle aux constituants ayant un rôle sémantique. Cette tâche est un problème de classification multi-classes. Les SVM sont des classificateurs binaires, donc pour assigner les bons rôles aux constituants, nous entraînons un classificateur pour chaque rôle. Lors de l'entraînement d'un classificateur pour un rôle, les constituants ayant ce rôle sont des exemples positifs, tandis que tous les autres sont des exemples négatifs. Ces classificateurs sont ensuite utilisés dans une configuration un contre tous (OVA) pour identifier le meilleur rôle à assigner au constituant.

Pour réaliser le repérage des rôles sémantiques nous avons fait plusieurs expériences, principalement sur la première étape de classification. La deuxième étape de classification ne peut pas effacer les erreurs d'identification de la première étape. Dans la seconde étape, le classificateur est entraîné seulement sur des constituants qui ont un rôle sémantique, ainsi, un constituant mal identifié dans la première étape se voit donc assigné un rôle dans la deuxième. Une telle erreur n'est pas catastrophique dans notre cas, puisque lors de la classification des constituants, le classificateur ne tient pas compte du fait qu'un rôle à déjà été assigné, ainsi, deux constituants peuvent avoir le même rôle pour le même prédicat.

Avant de pouvoir entraîner les classificateurs, nous devons transformer le problème dans un format compatible avec l'algorithme de classification. Dans un premier temps, nous convertissons les données du PropBank et du Penn TreeBank en un ensemble d'exemples. Pour l'entraînement du premier classificateur nous séparons les exemples en deux ensembles positifs et négatifs : les positifs ont un rôle sémantique, les négatifs n'en ont pas. Pour l'entraînement des classificateurs de la deuxième étape, nous prenons seulement les exemples positifs, ceux qui ont un rôle sémantique, que nous séparons encore en fonction du rôle qu'ils ont. Ainsi, pour chaque classificateur nous créons deux ensembles d'exemples : positifs et négatifs, en fonction du rôle sur lequel

nous entraînons le classificateur. Pour nos expériences, nous conservons la subdivision du Penn TreeBank en 25 parties, ceci nous permet d'évaluer la performance des méthodes sur différents sous-ensembles. Cette subdivision est essentielle pour le développement des méthodes, puisque l'extraction des attributs produit 754 000 attributs sur lesquels nous entraînons nos classificateurs. Cette explosion du nombre d'attribut est due à l'algorithme SVM que nous entraînons sur des attributs binaires (valeur de 0 ou 1).

6.3.3 Expériences et résultats

Les algorithmes d'apprentissages comme SVM dépendent énormément des paramètres utilisés lors de l'entraînement. Le paramètre le plus important est la fonction de noyau utilisée, $K(x, x')$ où $x, x' \in \mathbb{R}^n$ sont deux candidats, elle détermine la similarité entre les éléments. Lorsque $K(x, x') = \langle x, x' \rangle$ (le produit scalaire de x et x', la distance usuelle entre deux éléments) la fonction de noyau est linéaire. Dans le cas de la fonction de noyau linéaire, chaque attribut i du candidat x sera comparé à l'attribut i du candidat x'. Dans le cas des fonctions de noyau polynomiales, $K(x, x') = (\langle x, x' \rangle + l)^d$, elles nous permettent de considérer des d-uples d'attributs, tout en évitant l'explosion combinatoire. Dans nos expériences, nous avons considéré les fonctions de noyau polynomial de degré 2 et 3, ce qui nous permet de considérer les couples et les triplets d'attributs. Une fonction de noyau peut aussi être créée spécifiquement pour solutionner un problème [46]. Dans nos expériences, nous avons aussi utilisé la fonction de noyau gaussien RBF, $K(x, x') = \exp(-\frac{\|x-x'\|}{2\sigma^2})$, parce qu'elle est suggérée comme première fonction à essayer [30]. De plus, elle est non-linéaire, le nombre d'hyperparamètres qui l'influencent est moins grand qu'un noyau polynomial et elle est plus stable numériquement. Les autres paramètres des SVM déterminent la tolérance aux mauvaises classifications et le critère de convergence de l'algorithme lors de l'entraînement.

Classification des constituants

Les résultats que nous rapportons dans cette section sont le résultat d'une vingtaine d'expériences réalisées pour déterminer quelle implémentation de l'algorithme SVM est la plus efficace et quels paramètres nous devons utiliser pour entraîner un

modèle solutionnant notre problème. Pour certaines expériences, nous n'avons pas de résultats parce que : notre espace d'attributs est trop gros ou les implémentations de l'algorithme ne sont pas stables avec les données de notre problème.

Les résultats que nous obtenons en utilisant le logiciel *libsvm* [13] avec un noyau gaussien RBF ne permettent pas d'établir un point de départ d'où nous pourrons améliorer nos résultats. L'entraînement d'un classificateur avec *libsvm* nécessite, avec une seule section du corpus, de 5 à 6 heures de calcul, mais tous les candidats sont classifiés comme des constituants sans rôle sémantique par ce classificateur. Lorsque nous entraînons un classificateur sur 4 parties du corpus, le logiciel termine le calcul après 5 jours avec un message d'erreur, sans avoir produit de modèle. Cette augmentation du temps de calcul est *normal* parce que la complexité de calcul de l'algorithme d'entraînement est exponentielle en fonction du nombre d'exemples à entraîner.

Puisque le temps d'entraînement est fonction du nombre d'exemples, en éliminant une partie des candidats négatifs, pour n'en garder que 1/15, nous diminuons le nombre total d'exemples et nous équilibrons le nombre d'exemples positifs et négatifs. Ainsi, sur les parties 1 à 4 du corpus nous avons 75 706 exemples positifs et 97 357 exemples négatifs. En utilisant cette méthode, nous obtenons le meilleur résultat avec le logiciel SVM^{light} [31] et une fonction de noyau gaussien. Les tests sur la section 5 du corpus (avec 1/15 des exemples négatifs) donnent une *accuracy* de 91,5%, une précision de 91,9% et un rappel de 88,3%. L'*accuracy* est le nombre de candidats bien classifiés divisé par le nombre total de candidat. Lorsque nous testons ce modèle de classificateur sur la partie 22 du corpus, l'*accuracy* est de 94,18%, mais la précision et le rappel deviennent 0. L'élimination d'une partie des exemples négatifs n'est donc pas une bonne solution parce que l'algorithme n'a plus assez d'attributs associés aux exemples négatifs pour les séparer.

SVM^{light} est le logiciel que nous privilégions pour entraîner les modèles lors des expériences suivantes. La quantité d'attributs et d'exemples de notre problème ne cause pas de problème au logiciel et nous pouvons facilement modifier et optimiser le calcul de la fonction de noyau.

Pour pallier à la distribution inégale des exemples positifs et négatifs, nous modifions un facteur de coût qui augmente la conséquence d'une erreur sur un exemple positif lors de l'entraînement. En ajustant ce facteur de sorte qu'un exemple positif mal classifié a un poids 5 fois plus grand qu'un exemple négatif mal classifié, nous

augmentons le taux de rappel, mais nous faisons diminuer la précision. Un modèle entraîné sur la section 1 du corpus en utilisant un noyau polynomial de second degré $K(x, x') = (1 + \langle x, x' \rangle)^2$ donne sur les sections 5 et 6 du corpus des résultats similaires, une *accuracy* de 97,2% pour les deux, une précision/rappel de 76,9%/64,9% pour la section 5 et 77,1%/65,3% pour la section 6. En utilisant la même fonction de noyau, mais en ajustant le facteur de coût d'une mauvaise classification positive à une valeur de 5, l'*accuracy* pour les deux sections diminue à 96,4%, la précision à 59.4%, mais le rappel augmente à 91,6% et 92,2% respectivement.

L'entraînement d'un classificateur sur l'ensemble du corpus nécessite énormément de temps de calcul, mais pour évaluer l'applicabilité de cette solution à notre problème, nous n'avons pas besoin de tous les verbes. Nous prenons seulement 40 verbes extraits du corpus de courriels BCE-4 et toutes les phrases du Penn TreeBank où un de ces verbes apparaît. Nous obtenons ainsi 1 294 000 exemples en utilisant les phrases des sections 2 à 21 du Penn TreeBank, que nous utilisons pour entraîner un modèle avec une fonction de noyau polynomiale de degré 3, $K(x, x') = (1 + \langle x, x' \rangle)^3$, avec un facteur de coût d'erreur positive de 2. L'entraînement de ce modèle nécessite 68 heures, presque 3 jours de calcul. En utilisant les phrases contenant nos verbes de la section 23 du corpus, nous obtenons une *accuracy* de 96,8%, une précision de 66,1% et un rappel de 86,3%.

Les résultats que nous obtenons lors de cette première étape de classification sont proches de ceux obtenus avec les systèmes utilisant un modèle de classification entraîné sur un corpus complet. La précision reste plus faible que ce qui est rencontré dans la littérature, par contre ces exemples faussement identifiés positivement pourront être éliminés lors de la classification des rôles sémantiques, c'est-à-dire au moment du choix du meilleur rôle sémantique d'un constituant.

Classification des rôles sémantiques

Dans la première étape de classification, nous avons un problème avec la quantité de données à traiter, qui n'apparaît pas lorsqu'on veut assigner un rôle à un constituant. Dans cette étape, le modèle de classification est entraîné seulement sur les constituants qui ont un rôle sémantique. La méthode que nous utilisons pour classifier les rôles consiste à entraîner 5 classificateurs, où chaque rôle est mis en com-

pétition avec les autres sous une configuration un contre tous (*one-vs-all* ou OVA). Nous avons privilégié cette approche puisqu'elle performe généralement aussi bien que les séries de duels (AVA) [19, 59] et nous avons alors moins de modèles à entraîner.

Les 5 classificateurs que nous entraînons sur les sections 2 à 22 du Penn TreeBank séparent les arguments : ARG0, ARG1, ARG2, ARG3 et ARGM. Les rôles ARG4 et ARG5 ne sont pas considérés pour l'entraînement de classificateurs parce qu'ils ne sont pas assez fréquents et tiennent plusieurs rôles sémantiques différents, ce qui diminue la corrélation entre ces rôles et les attributs. Les classificateurs sont tous entraînés de la même façon avec SVM^{light}, une fonction de noyau polynomiale de degré 2 et les paramètres standards.

Les classificateurs donnent de bons résultats pour la classification des rôles sémantiques (tab. 6.2) sur la section 23 du Penn TreeBank. En utilisant la stratégie un contre tous, nous prenons, pour chaque exemple, le classificateur qui lui donne la plus grande valeur positive, ce qui explique la présence d'une colonne *null* dans les résultats. La *colonne* total indique le nombre total de rôles présents dans le corpus de test et la *ligne* total indique le nombre de constituants identifiés par le classificateur. La précision et le rappel de chacun des classificateurs, dans le contexte de la configuration un contre tous, est rapporté dans la dernière ligne et la dernière colonne. Si nous obligeons tous les candidats à avoir un rôle, la précision diminue, mais le rappel augmente, particulièrement pour les rôles ARG2 et ARG3. L'*accuracy* du classificateur est de 92,1%.

	ARG0	ARG1	ARG2	ARG3	ARGM	null	Total	Rappel
ARG0	6653	98	6	0	19	36	6812	97,7
ARG1	221	5285	36	7	60	149	5758	91,8
ARG2	18	149	1060	0	144	218	1589	66,7
ARG3	0	15	9	108	33	69	234	46,2
ARG4	0	1	14	0	11	126	152	-
ARG5	0	0	0	0	4	13	17	-
ARGM	17	44	68	3	7640	186	7958	96,0
Total	6909	5592	1193	118	7911	797	22520	
Précision	96,3	94,5	88,9	91,5	96,6	-		92,1

TAB. 6.2 – Résultats de la classification des rôles sémantiques sur la section 23 du Penn TreeBank

6.4 Conclusion

Les résultats du repérage des rôles sémantiques en utilisant la méthode d'apprentissage statistique supervisée des machines à vecteur de support (SVM), sont pertinents pour analyser les documents et les questions. La classification des constituants ayant un rôle sémantique pourrait être améliorée légèrement par une sélection plus précise des attributs influençant la tâche, même si nos résultats s'approchent des résultats rapportés par les autres équipes de recherche travaillant sur ce problème. Les résultats obtenus lors des expériences par notre implémentation s'expliquent par les décisions prises lors de la préparation des données et de la réalisation de la méthode. Les modèles de classificateurs, entraînés lors des expériences, sont assez polyvalents pour être utilisés avec des données d'autres domaines et les attributs que nous utilisons peuvent être extraits avec un analyseur syntaxique et un identificateur de mots clés. La polyvalence de nos modèles empêche l'optimisation de la représentation de l'espace d'attributs, ce qui a pour conséquence d'augmenter le temps requis pour entraîner un modèle et de diminuer leur performance, puisqu'ils ne peuvent être entraînés sur l'ensemble des données. Dans un cadre moins expérimental, en figeant les attributs considérés, nous pourrions en compresser la représentation pour diminuer le nombre d'attributs que l'algorithme SVM devra tenir compte.

La répartition inégale des exemples positifs et négatifs pour l'entraînement des classificateurs diminue la qualité de la classification réalisée par une SVM. Cette caractéristique du problème doit être considérée pour améliorer les résultats de la classification [1, 14]. Ces facteurs de performances sont liés au contexte pour lequel les expériences sont réalisées, les meilleurs résultats sont rapportés dans un contexte expérimental, dans notre cas, les résultats proviennent d'une architecture pour repérer les rôles sémantiques dans un contexte d'utilisation normale. L'amélioration des résultats de cette étape est prématurée dans le cadre de notre architecture.

Le but d'implémenter un analyseur sémantique pour repérer les structures prédicat-arguments est d'associer les actions et les entités des textes avec celles mentionnées dans les questions. Par le repérage des structures prédicat-arguments, nous voulons aussi être en mesure d'analyser les questions pour déterminer la fonction et la forme de la réponse attendue. Ainsi, certaines combinaisons de prédicat-arguments pourraient être associées à un traitement spécifique pour répondre à la question.

L'analyseur sémantique doit être performant, puisqu'il sert à identifier un rôle manquant dans une question, qui sera recherchée dans le corpus de documents [47]. Mais le repérage des rôles sémantiques pour les questions contenues dans les courriels du corpus BCE-4 ne fonctionne pas très bien avec les modèles de classification entraînés sur le Penn TreeBank. Les mauvais résultats de l'analyse sémantique sur les questions sont attribuables principalement à la spécificité du problème. La syntaxe des questions est différente des phrases du Penn TreeBank et puisque nous utilisons l'analyseur syntaxique de Stanford, utilisant une grammaire probabiliste entraînée sur le Penn TreeBank, les analyses syntaxiques des questions doivent être modifiées manuellement pour qu'elles soient valides. Même avec des analyses syntaxiques correctes des questions, le repérage des rôles sémantiques est encore problématique parce que les attributs sur lesquels nous avons entraîné les modèles de classificateur ne sont pas liés de la même façon avec le rôle des arguments dans les questions et dans le Penn TreeBank.

La disponibilité de données adaptées à la syntaxe des questions et aux domaines d'application du système rendrait cette méthode beaucoup plus intéressante à exploiter. Même si les résultats obtenus nous indiquent que le système doit être amélioré, l'entraînement sur un ensemble de données plus près du problème que nous traitons est une solution plus adaptée à notre problème, qui ne passe pas par le développement fastidieux d'un système à base de règles. En combinant le traitement des rôles sémantiques avec des connaissances sur les entités et les concepts rencontrés dans notre domaine, l'extraction d'une réponse à une question devient une étape plus formelle, qui se distingue des méthodes où la réponse est extraite à partir d'un segment de texte.

Chapitre 7

Identification des réponses

Notre solution au problème de la réponse automatisée aux courriels est fondée sur l'hypothèse que nous pouvons le traiter comme un problème de question-réponse, mais dans lequel nous devons considérer les difficultés de traitement liées à la nature des courriels et à la spécificité du domaine du problème. En question-réponse factuelle, les réponses sont récupérées à partir d'un corpus de textes journalistiques, une situation différente de celle rencontrée dans notre problème. La question d'un utilisateur consultant le service d'information est généralement plus complexe qu'une question factuelle. L'utilisateur a probablement déjà effectué une recherche à partir du site web de l'entreprise et il n'a pas trouvé de réponse à sa question. Il est aussi possible que l'utilisateur recherche une information concernant : un événement futur, le traitement d'un dossier ou la confirmation d'une inscription. En comparaison, la difficulté des questions factuelles des conférences TREC est beaucoup moins élevée. Les premières conférences TREC où la question-réponse fût abordée ont démontré qu'un engin de recherche utilisant l'extraction de passages était capable de retourner un passage contenant la bonne réponse avec précision. Un utilisateur disposant d'un engin de recherche et d'un corpus de référence contenant la réponse peut donc répondre à ses questions sans l'aide d'un préposé et avec un effort minimal. En considérant cette situation, il devient évident que nous devons exploiter plusieurs approches pour récupérer la réponse si nous voulons retourner automatiquement une information satisfaisant les besoins des utilisateurs.

Dans le cadre de notre projet, nous avons été confrontés à un problème d'accès aux données. Les ressources dont nous avons besoin pour répondre aux courriels n'étaient pas accessibles et les réponses aux courriels ne sont généralement pas contenues dans les données textuelles provenant du site web de l'entreprise. L'information que nous possédons pour réponse aux courriels est incomplète, ainsi, notre approche pour identifier la réponse ne peut être comparée avec ce qui est réalisé lors de TREC. Puisque nous ne pouvons pas répondre correctement aux courriels à partir des sources d'informations disponibles, il nous est impossible d'évaluer précisément notre approche pour l'extraction de la réponse.

7.1 Exploitation des sources d'information

Lorsque nous avons analysé les courriels du corpus BCE-4 nous avons identifié des propriétés nous permettant de classifier les questions en fonction du traitement qu'une personne doit réaliser pour y répondre. Puisqu'il y a plusieurs taxonomies de questions, il y a aussi plusieurs méthodes pour trouver les réponses. Notre approche est donc de considérer les ressources qui sont nécessaires pour répondre aux questions et d'explorer des manières d'exploiter ces ressources efficacement.

Dans le chapitre 4, nous avons donné plusieurs exemples des questions provenant du corpus BCE-4. Nous avons pu y observer que la réponse aux questions nécessitait une bonne connaissance du domaine des relations aux investisseurs et une capacité d'analyse des procédés d'affaires pour répondre aux questions concernant des processus en cours de traitement. Pour intégrer les connaissances nécessaires pour répondre aux courriels, nous avons réalisé une modélisation du domaine en fonction des sources disponibles et des questions du corpus BCE-4. Lors de cette étape, nous avons constaté qu'il était difficile de récupérer l'information à caractère fonctionnelle (p. ex. les formules de calcul et les démarches nécessitant plusieurs étapes) qui permet à l'utilisateur de réaliser une tâche particulière.

7.1.1 Représentation formelle du domaine

Les ressources permettant de répondre aux courriels proviennent de plusieurs domaines, c'est pourquoi nous devons exploiter plusieurs niveaux de formalisation

de l'information : des connaissances générales aux particularités techniques du domaine. En question-réponse factuelle, le prétraitement des documents permet d'extraire l'information pertinente à partir des données textuelles, dans notre problème, l'information nécessaire pour trouver des réponses ne peut pas être repérée de façon satisfaisante dans les textes. Conséquemment, l'approche que nous préconisons pour représenter les connaissances exploitées dans ce problème est de modéliser le domaine de référence par une ontologie combinant les connaissances des domaines touchés par les relations aux investisseurs et nécessaires pour répondre aux courriels.

L'ontologie utilisée pour modéliser les connaissances constitue un cadre dans lequel nous représentons l'information. Ce cadre de représentation nous permet d'exploiter l'information comme le schéma d'une base de données. Notre représentation du domaine ne se limite pas à celui d'une base de données, la représentation des connaissances par une ontologie doit nous permettre d'associer les requêtes aux bonnes ressources et les ressources aux concepts y correspondant. Cette façon d'aborder le problème est essentiellement la même que celle des premiers systèmes de question-réponse utilisant une base de donnée pour trouver la réponse à une question traduite sous la forme d'une requête. La différence consiste à ce que les connaissances que nous devons formaliser sont disparates, certaines sont éminemment plus complexes à représenter et nous voulons une solution souple pouvant être réutilisée dans un autre domaine que celui des relations avec les investisseurs.

La modélisation du domaine des services de relations avec les investisseurs nécessite une représentation possédant un pouvoir expressif nous permettant de représenter plusieurs types d'entités et de relations. Ainsi, nous devons exploiter des représentations plus expressive qu'un langage contrôlé ou un thésaurus. L'ontologie est une représentation plus adaptée à nos besoins parce qu'elle nous permet de lier les instances et les concepts entres eux.

Nous avons conçu l'ontologie selon une approche ascendante ; nous avons débuté la conception en énumérant les entités et les relations faisant l'objet d'une question dans le corpus BCE-4, pour ensuite, lier les entités en utilisant les relations pertinentes à notre problème. Cette première étape nous a permis d'établir la liste des concepts propres au domaine, que nous avons ensuite regroupés par thèmes pour identifier plus facilement les relations les unissant. Dans le cas présent, nous n'avons pu établir précisément un lexique du domaine, puisque la liste des concepts provient

principalement des questions contenues dans les courriels, mais la réponse n'y figure évidemment pas. De plus, comme nous le mentionnions lors de l'étude des questions du corpus, nous avons un problème au niveau de l'interprétation sémantique des termes des courriels parce que les utilisateurs n'ont pas tous la même compréhension du discours du domaine. Ceci diminue l'apport de notre ontologie pour le traitement des courriels, car l'ontologie ne pourra pas être utilisée pour traiter la requête du courriel sans avoir désambiguïsé chaque mot en lui assignant un seul sens.

La modélisation des entités associées aux *concepts de base* pour notre problème ne comporte pas de difficulté particulière. Nous avons défini les concepts en fonction de leurs attributs et des relations avec les autres entités, dans le but d'avoir une représentation autonome. Dans un premier temps, nous avons identifié et modélisé les concepts suivants :

- Person
- Company
- Stock
- Financial Instrument
- Share
- Bond
- Dividend
- Date
- Currency Measure

Nous avons ensuite identifié les relations liant les concepts et les entités. La définition de ces relations est essentielle pour extraire les réponses aux questions portant sur une entité en relation avec une autre. Nous avons modélisé un petit sous-ensemble des relations nécessaires à la représentation des connaissances du domaine. Nous savons par l'expérience acquise lors du projet de création du langage TimeML, que la modélisation des relations entre les entités peut rendre le modèle incompréhensible si nous n'avons pas une vue globale sur le domaine. Parmi l'ensemble des relations liées au domaine d'application, nous avons seulement considéré les relations pouvant constituer un modèle intéressant du domaine et permettre d'extraire des réponses aux questions de notre corpus :

- stockOf(Stock, Company)
- stockHolder(Stock, Person)
- issuedBy(FinancialInstrument, Agent)
- convertibleIn(Stock1, Stock2)
- convertibleDate(Stock, Date)
- maturityDate(FinancialAccount, Date)
- splitFor(StockSplit, CurrencyMeasure1, CurrencyMeasure2)
- closingPrice(Stock, CurrencyMeasure, Date)

- dividend(Stock, CurrencyMeasure, TimeInterval)
- employee(Person, Org, Pos)

Certaines de ces relations pourrons ensuite être extraites des questions et des textes en utilisant la méthode de repérage des rôles sémantiques du chapitre précédent. Nous pourrons aussi exploiter une partie des données structurées de l'entreprise pour compléter la base de connaissances fondée sur cette ontologie.

Nous avons développé l'ontologie avec l'outil de création d'ontologie Protégé[1] et en utilisant le formalisme de représentation OWL-DL. Lors du développement de l'ontologie, nous avons rencontré un problème d'expressivité du langage parce que le formalisme OWL-DL ne nous permet pas de représenter les connaissances de notre domaine pour exploiter les engins de raisonnement adaptés à OWL. Un problème que nous rencontrons concerne la modélisation des relations. Lors de la définition d'une relation avec OWL-DL, nous pouvons définir un domaine et une image, par contre, cette définition n'est pas considérée comme une contrainte, mais comme une caractéristique définissant une instance de la classe de l'image ou du domaine. Pour exploiter notre modèle nous avons besoin d'une représentation qui puisse exploiter des relations et des descriptions de concept sous la forme de règles.

Pour solutionner notre problème, la modélisation du domaine avec une ontologie n'est pas la solution à tous les maux. Ce type de représentation n'est pas adapté à l'information que nous devons manipuler pour récupérer efficacement la réponse. Pour solutionner notre problème sur les questions portant sur le contenu informatif, il nous faut encoder manuellement l'ensemble des entités et des relations, l'extraction automatique de l'information ne s'applique pas à notre modélisation. La modélisation des connaissances de ce problème est difficile, l'ingénierie des connaissances doit être réalisée par un expert du domaine. Nous avons abordé le problème avec nos connaissances des domaines de l'investissement, des finances et des corporations, qui ne sont pas celles d'un expert. Nous avons quand même modélisé convenablement certaines parties du domaine, mais nous nous sommes rendu compte qu'il y a de nombreuses différences dans les pratiques et la terminologies des entreprises.

Nous avons donc écarté l'utilisation d'une modélisation du domaine par une ontologie pour solutionner spécifiquement notre problème. Par contre, nous croyons qu'avec les ressources nécessaires à la définition du domaine disponible et l'utilisation

[1] `http://protege.stanford.edu/`

d'un modèle de représentation plus expressif que le langage d'encodage d'ontologie OWL, le problème de l'extraction de la réponse est grandement facilité. Dans notre problème, nous n'avons pas exploré en détail les autres types de représentations qui pourraient solutionner notre problème. Sans posséder le même courant de sympathie que les ontologies, nous avons exploré la représentation logique en clause de Horn pour traiter les données concernant les actions. Nous avons implémenté cette modélisation comme un programme logique Prolog, cette solution nous permet de typer les relations et de définir des critères d'appartenance aux classes d'objet. Aussi, nous pouvons encoder et réaliser une partie des calculs qui sont demandés dans les questions, ce qui ne pouvait être réalisé avec OWL. L'utilisation d'une telle représentation est envisageable pour implémenter un système de réponse aux courriels. Toutefois, cette solution n'est pas très économique, car elle demande plus de ressources et dédoublerait ce qui se fait déjà dans l'organisation au niveau des services d'information, sans pouvoir les remplacer, car l'information doit rester facilement accessible pour les personnes ayant besoin de l'exploiter.

7.2 Conclusion

L'identification de la réponse est essentielle pour répondre automatiquement aux courriels. Dans notre problème, nous ne pouvons pas exploiter les mêmes techniques qu'en question-réponse factuelle parce que la nature des données interrogées n'est pas la même. Puisque nous ne pouvons pas exploiter efficacement les données textuelles, nous avons exploré l'avenue de la représentation structurée des données. Nous obtenons des résultats mitigés avec la représentation de l'information sous la forme d'une ontologie ; la modélisation formelle du domaine semble être une piste prometteuse, mais nous n'avons pas les données, les ressources et les compétences pour créer une modélisation exhaustive des connaissances pour solutionner notre problème. La spécificité de l'information faisant l'objet des requêtes dans les courriels fait en sorte que l'identification de la réponse devient une tâche difficile ; un système répondant à toutes les requêtes devrait être composé d'une multitude de systèmes experts spécialisés.

Chapitre 8

Prétraitement des documents de référence

Le prétraitement des documents de référence sert à identifier l'information générale susceptible d'être recherchée ou utilisée pour répondre à la question à partir d'un corpus de documents. En question-réponse factuelle, le prétraitement des documents de référence consiste généralement à extraire certaines entités nommées et les relations définissant des rôles sémantiques simples. L'extraction de l'information est réalisée par des patrons prédéfinis ou des algorithmes statistiques, mais la portée de ses méthodes reste limitée parce que la rédaction manuelle de patrons est fastidieuse et les algorithmes statistiques ont besoin de données où l'information recherchée est annotée. Le développement des méthodes de prétraitement des documents spécifiquement pour la question-réponse a été négligé, car plusieurs systèmes de question-réponse utilisent les engins de recherche et l'information d'internet pour récupérer les documents pertinents.

Dans l'automatisation de la réponse aux courriels, l'hypothèse de la redondance de l'information n'est pas exploitable comme en question-réponse factuelle. Le prétraitement des documents de référence doit donc identifié toute l'information dont nous avons besoin. L'information recherchée peut n'apparaître que dans un seul document, nous ne pouvons passer à côté de cette information. Nous abordons cette étape en considérant l'aspect temporel de l'information : les liens entre le temps et les événements. Le temps a un rôle important dans le domaine des investisseurs, la succession

des événements entraîne la péremption de l'information. Par l'ajout de repères temporels à l'information, nous pouvons établir explicitement les relations temporelles entre les événements. Nous considérons donc l'ajout de balises d'annotation TimeML pour le prétraitement des documents.

8.1 Annotation temporelle des documents sources

L'information temporelle est présente dans tous les documents rapportant de l'information sur une situation, le système d'information que nous développons exploite ce type de documents. Pour identifier les relations entre les faits et les événements, nous devons reconnaître les marqueurs temporels et les événements rapportés dans les documents de référence. Dans les articles de nouvelles, ces éléments rapportent des événements se développant dans le temps, ayant pour conséquence de modifier l'état du monde auquel on fait référence. Les marqueurs temporels permettent au lecteur d'établir les points marquants d'une histoire. Les événements sont rarement ancrés de façon absolue dans le temps, plutôt, nous interprétons l'ordonnancement des événements par notre capacité de raisonner sur les relations temporelles sous-jacentes au texte.

Nous retrouvons des expressions temporelles dans les textes informatifs, ces expressions peuvent être catégorisées de la façon suivante :

- Les **expressions explicites** sont les plus évidentes, elles peuvent être positionnées de façon unique sur une ligne du temps. P. ex., *juin 2005, 20:15 le 18 avril 2001.*
- Les **expressions déictiques** sont exprimées en relation avec le moment de la rédaction du document. P. ex., *demain, l'an passé, dans deux semaines.*
- Les **expressions relatives** sont en relation avec le temps relaté dans le récit correspondant au moment d'un événement. P. ex., *durant la manoeuvre, après le ..., le jeudi.*
- Les **expressions de durées** font référence à un intervalle de temps. P. ex., *pendant deux jours.*
- Les **expressions récurrentes** font référence à des événements qui se répètent de façon périodique. P. ex., *chaque lundi.*

- Les **expressions d'occurrences** font référence aux événements qui se produisent un certain nombre de fois sur une période de temps déterminée. P. ex., *deux fois par semaine, quatre examens hier*

L'identification des expressions temporelles explicites dans les textes journalistiques peut être réalisée automatiquement par des algorithmes à base de patrons [42]. En question-réponse et dans notre problème, nous devons considérer les composantes temporelles présentes dans plusieurs questions pour trouver la réponse. Nous pouvons identifier automatiquement les expressions temporelles explicites, mais plusieurs questions nécessitent l'ancrage absolu des événements dans le temps et la compréhension des relations de cause à effet. À titre d'exemple, prenons l'extrait suivant provenant d'un article de presse[1] provenant du corpus TimeBank[2] de documents annotés par TimeML :

> *Turkey (AP) - Some 1,500 ethnic Albanians marched Sunday in downtown Istanbul burning Serbian flags to protest the killings of ethnic Albanians by Serb police in southern Serb Kosovo province. The police barred the crowd from reaching the Yugoslavian consulate in downtown Istanbul, but allowed them to demonstrate on nearby streets.* (tronqué)

Cet extrait contient une suite d'événements qui sont liés entre eux. Par exemple, l'expression déictique *Sunday* fait référence à l'expression temporelle explicite du 2 août 1998, parce que l'événement *marched* est au passé et nous avons la date de rédaction de la nouvelle, le 3 août 1998, dans l'en-tête de l'article. Les questions que nous voulons traiter font appel à la compréhension de la structure temporelle de la nouvelle, cette structure peut être identifiée à partir des marqueurs temporels du texte. Les question suivantes nécessitent la compréhension des relations temporelles et peuvent être répondues à partir de l'extrait d'article précédent :

1. What happened August 2nd 1998 ?
2. Were Serbian flags burned before the killings ?
3. Were ethnics Albanians killed during the demonstration ?

La question 1 fait appel à un raisonnement simple où il faut calculer des dates à partir d'un calendrier et ensuite extraire l'action du texte. La question 2 demande d'être capable de comprendre que la relation entre la *protestation* et l'événement

[1] Time Bank APW19980308.0201 (Associated Press newswire)

[2] Une description sommaire du corpus se trouve à la p. 134

passé est temporelle et que l'événement *burning Serbian flags* a eu lieu après ou durant l'événement *killings of ethnic Albanians*. La question 3 est plus difficile, car elle demande un ordonnancement d'événements qui n'apparaissent pas dans la même phrase. La relation entre les deux événements peut être établie par transitivité de la relation d'ordre chronologique sur l'ensemble des événements de l'histoire.

Ces questions mettent en évidence certains problèmes que nous devons considérer pour répondre aux questions, même les plus simples, à propos du temps et des événements. Nous pouvons identifier deux types de modèles du temps, le modèle physique et le modèle linguistique. Le modèle de représentation du temps le plus commun est le modèle physique qui considère le temps comme un point sur une ligne. C'est un modèle pratique pour modéliser le monde par des équations mathématiques, mais d'une utilité très limitée pour comprendre les communications. Les modèles linguistiques sont utiles pour analyser le langage, mais ils sont difficiles à interpréter lorsque nous voulons les exploiter dans le cadre d'une représentation formelle du temps.

Puisqu'il y a plusieurs façons d'analyser et de représenter la temporalité des événements, nous pouvons commencer par étiqueter l'information temporelle que nous voulons analyser. Par la suite, nous pourrons spécifier le cadre d'analyse, le but étant d'exploiter les relations temporelles pour une application et un domaine précis. L'étiquetage des expressions temporelles permet d'identifier les événements et, évidemment, les expressions temporelles, de sorte que nous puissions établir des relations entre eux. Dans l'exemple précédent, l'étiquetage permet de spécifier que l'événement *marched* a lieu *Sunday* et que l'événement *burning Serbian flags* est arrivé après *the killing of ethnics*, mais avant la démonstration (*to demonstrate*). Cette annotation permet d'établir un ordre des événements à partir des relations que l'annotateur a étiquetées.

8.1.1 TimeML

Le langage d'annotation TimeML a été développé pour étiqueter l'information temporelle dans les textes dans le but de répondre aux questions concernant les événements et les entités temporelles. L'annotation de l'information temporelle directement dans le texte avec TimeML permet d'identifier les représentations temporelles et les compléments d'information pouvant déjà être annotés dans le texte. Nous avons

participé au développement de TimeML dans le but de solutionner quatre problèmes liés à l'identification des événements et de la temporalité :

1. l'ancrage absolu des événements dans le temps ;
2. l'ordonnancement chronologique et relatif des événements entre eux ;
3. le raisonnement à partir d'expressions temporelles sous-spécifiées dans le contexte d'énonciation ;
4. le raisonnement sur la persistance des événements.

Le langage TimeML est construit autour des concepts d'objet et de relation temporels. Les objets temporels sont les événements et les expressions temporelles explicites. Les relations temporelles sont identifiées par des *liens* et divisées en trois classes : liens temporels, liens subordonnants, liens aspectuels. Nous décrivons plus en détail la nature de ces liens et le langage TimeML dans l'annexe A.1.

8.1.2 Annotation des documents

Nous pouvons réaliser l'annotation des étiquettes identifiant les événements (`<EVENT>`) et les marqueurs temporels (`<TIMEX3>`) en utilisant des règles pour identifier les principaux événements et marqueurs temporels du texte. Par contre, nous n'avons pas encore de méthode pour annoter les liens entre les entités temporelles, car TimeML est encore un formalisme en développement et la quantité de documents annotés avec les balises de TimeML est limitée. L'annotation de documents avec TimeML est laborieuse, parce que le langage est dense et l'annotateur doit retenir plusieurs informations concernant les événements faisant référence à diverses parties du texte. Et comme nous le voyons dans l'exemple de la page 137, la réalisation manuelle de l'annotation nécessite énormément de gestion d'identificateurs, nécessitant parfois plusieurs étapes pour déterminer l'identificateur source. L'utilisation d'outils d'annotation efficaces est essentielle pour produire des annotations de qualités.

Un outil d'annotation doit surmonter les difficultés de l'annotation qui sont caractérisées par :

1. une annotation dense,
2. une vitesse d'annotation lente,

3. des inconsistances difficiles à éviter,
4. un faible taux de correspondance entre les annotateurs (*inter-annotator agreement*).

La haute densité de l'annotation provient du fait que le nombre de relations possibles entre les objets temporels, qui sont eux-mêmes très denses, croît de manière quadratique en fonction du nombre d'objets [66]. Lorsque la longueur du document augmente, le nombre d'événements augmente et le nombre de relations possibles devient rapidement difficile à gérer par un annotateur, sans l'aide d'outils d'automatisation. À titre d'exemple, les documents du corpus de documents annotés par TimeML, TimeBank, contiennent en moyenne 51 objets temporels (`<EVENT>` et `<TIMEX3>`) et peuvent être reliés entre eux de milliers de façons, car un objet temporel peut être mis en relation avec plusieurs objets. Ces relations, logiquement, ne sont pas toutes présentes dans le document, les documents du TimeBank contiennent en moyenne 48 liens temporels annotés. Mais en réalité, il y a plus de liens temporels que cela dans les documents, parce que les annotateurs n'annotent pas nécessairement les relations réciproques et la fermeture transitive des relations.

L'annotation d'un document prend beaucoup de temps en comparaison à l'étiquetage des entités nommées ou des étiquettes grammaticales, qui peut être réalisé en suivant le texte sans revenir en arrière. L'annotation des liens temporels nécessite la spécification d'un couple d'objets temporels et la spécification de différents attributs. De plus, les objets temporels ne se trouvent pas toujours près l'un de l'autre dans le texte.

L'examen de documents annotés manuellement nous permet de constater qu'il est difficile d'annoter un document sans introduire d'inconsistance. Les inconsistances proviennent de l'ambiguïté des relations temporelles dans le texte. Par exemple, un annotateur décide que l'événement A précède l'événement B et que l'événement B précède l'événement C. Par transitivité nous concluons que A précède C, mais il est possible que l'annotateur en décide autrement et que A soit inclus dans C à cause de l'information disponible à ce moment dans le texte. Dès lors, l'ensemble des relations entre A, B et C doit être reconsidéré, car la représentation des événements devient ambiguë, mais les annotateurs ne le font pratiquement jamais. Ce genre d'inconsistance dans l'annotation est attribuable à la complexité de la tâche et elle se produit habituellement à la fin de l'article lorsque l'annotateur est moins concentré sur la tâche.

Le faible accord entre les annotateurs est largement attribuable à la faible proportion de liens par rapport à l'ensemble des liens possibles et à la combinaison de liens et d'attributs possibles. Lorsque les objets temporels ne sont pas déjà annotés, le problème d'accord entre les annotateurs est encore plus évident, puisque les objets n'étant pas les mêmes, les liens deviennent plus difficiles à comparer.

La nécessité d'avoir un outil adapté à la tâche d'annotation de document TimeML est bel et bien réelle, la représentation graphique des éléments temporels d'un texte est une approche pour faciliter le travail des annotateurs en leur permettant d'avoir une vue globale de l'information annotée. La combinaison d'une représentation graphique de l'annotation et des algorithmes de traitement automatique (fig. A.7, p.146) permet d'avoir un outil pour ajouter des annotations qui auraient été oubliées, d'accélérer le processus d'annotation d'un texte, d'éviter les inconsistances de l'annotation et, ainsi, d'améliorer la qualité de l'annotation d'un annotateur à l'autre.

Un outil utilisé pour annoter des documents TimeML est *Alembic Workbench* [18] qui permet d'ajouter facilement les étiquettes `<TIMEX3>` et `<EVENT>`, car cette tâche est réalisée linéairement. L'annotation des liens est réalisable en utilisant un tableau permettant de voir l'ensemble des liens d'un certain type. L'utilisation de l'interface tabulaire pour l'annotation des liens demande une compréhension intégrale des spécifications d'annotation et une certaine expertise avec Alembic. En combinant Alembic avec une interface graphique représentant l'information temporelle, il est possible d'accélérer l'annotation des documents et d'en améliorer la qualité.

8.1.3 Annotation graphique de TimeML

Nous avons développé l'outil d'annotation graphique TANGO (TimeML Annotation Graphical Organizer) parce qu'il est nécessaire d'aider les annotateurs en leur donnant une perspective graphique de la tâche d'annotation. TANGO est la combinaison de deux aspects essentiels pour l'annotation de documents avec TimeML : l'annotation semi-automatique et la visualisation des documents. L'annotation semi-automatique des liens consiste à calculer la clôture transitive et l'inverse des relations en ajoutant des liens, ceci produit la clôture temporelle [66].

Nous avons développé la représentation graphique de l'annotation d'un document en deux étapes. Dans le cadre de TERQAS, nous avons développé une méthode qui

nous permet de visualiser le résultat de l'annotation d'un document. Ce prototype nous a servi pour développer l'outil TANGO qui intègre aussi des algorithmes pour ajouter automatiquement les liens temporels de la fermeture.

Dans le modèle de représentation du temps de TimeML, le temps n'est pas considéré dans un sens strictement physique, il y est difficile d'associer un temps absolu aux événements pour les ancrer sur une ligne du temps. Le paradigme de la ligne du temps est restrictif pour la visualisation de certaines relations du langage TimeML qui ne peuvent être inscrites sur la ligne du temps, aussi, les relations n'ont pas toujours d'interprétation géométrique précise. De plus, certains événements se comparent difficilement, parce qu'ils ne sont pas réalisés dans la même échelle de temps. Dans l'exemple concernant les événements d'Istanbul, les événements *marched*, *protest*, *burning*, *barred*, *reaching* se produisent sur une échelle de temps comparable, mais l'événement *killing* ne peut se comparer visuellement aux autres, puisqu'il est beaucoup plus distant des autres événements que les autres événements entre eux. Ainsi, nous ne tenons pas compte de l'aspect géométrique du temps, parce que le temps est difficile à mesurer et à représenter fidèlement de façon linéaire, mais nous utilisons la topologie des événements induite par les liens temporels, aspectuels et subordonnants.

Notre prototype permettant de générer des représentations graphiques est issu du développement du langage d'annotation. Lors du développement du langage, nous devions régulièrement réfléchir à partir de diagrammes pour résoudre les difficultés de représentation du temps. Puisque nous développions un langage d'annotation avec comme toile de fond le modèle du temps, il devenait intéressant de formaliser ceci en transformant automatiquement l'annotation dans une représentation graphique basée sur les graphes, car nous avons des entités (noeuds) et des lien (arcs).

La tâche du prototype est de traiter un document TimeML et d'en extraire un graphe qui est ensuite affiché à l'annotateur. Le développement du prototype nous a permis de comparer l'interprétation visuelle au sens que nous voulions donner à chaque relation. Il est important que l'interprétation visuelle de l'annotation soit identique ou similaire à l'interprétation sémantique, ceci facilite la tâche des annotateurs, car ils ont alors une référence visuelle.

Nous avons utilisé le prototype d'outil de visualisation pour vérifier la validité des annotations en comparant l'annotation avec l'interprétation visuelle, réalisée menta-

lement, des événements d'un récit. Nous l'avons aussi utilisé pour clarifier certains points du langage qui n'étaient pas spécifiés correctement et qui pouvaient causer des problèmes pour la consistance de la représentation temporelle. La description que nous faisons du prototype comporte quelques irrégularités par rapport à la version courante (v.1.2.1) de TimeML, car le prototype n'a pas été adapté et il a été remplacé par l'outil TANGO.

La représentation graphique du court article de la page 140 (illustrée dans les figures A.4, A.5 et A.6, pages 142–144) met en évidence la pertinence de pouvoir compter sur la visualisation de l'annotation pour détecter les relations oubliées et ordonner temporellement les événements de gauche à droite.

La représentation graphique d'une annotation telle que nous l'avons réalisée lors de TERQAS est utile pour l'annotateur, mais nous pouvons la rendre plus efficace en intégrant cette visualisation à l'intérieur d'un outil d'annotation. Les outils d'annotation existants ne permettent pas d'annoter convenablement les liens temporels des documents. Nous avons donc développé l'outil d'annotation TANGO qui solutionne les problèmes de l'annotation des liens temporels. L'outil d'annotation TANGO est décrit dans l'annexe A.3.

8.2 Conclusion

L'utilisation du langage TimeML pour annoter l'information temporelle est pertinente pour notre problème. Les communiqués de presse émis par une compagnie sont similaires aux articles de presse, qui ont souvent comme source d'information ces communiqués. Dans le communiqué de presse suivant provenant du corpus de référence de BCE, nous pouvons constater la présence d'événements liés temporellement.

> **BCE Announces the Closing of the Sale of Excel**
> *MONTREAL, April 8 – BCE Inc. today announced the closing of the sale by Teleglobe Inc. of Excel Communications' North American operations to VarTec Telecom Inc. Final regulatory and other approvals have been received. The final proceeds of disposition were US $227.5 million, which have been paid in the form of five-year interest-bearing promissory notes.*

Les événements de cet exemple sont difficiles à interpréter hors du contexte global des événements de la compagnie BCE. Dans cet exemple, nous avons la date absolue de l'*annonce* à partir de l'en-tête du communiqué, mais nous ne pouvons pas déterminer l'ordre des événements à partir de l'information textuelle. L'information temporelle ajoutée par TimeML aux communiqués de presse pourra être exploitée dans le cadre de la recherche d'une réponse lorsque l'information temporelle aura été ajoutée à tous les documents de la collection. D'ici là, l'information temporelle recueillie à partir de quelques documents demeure trop éparse pour être exploitable.

Nous n'avons malheureusement pas pu exploiter TimeML directement dans nos travaux sur la réponse automatisée aux courriels, parce que la tâche d'annoter notre corpus de communiqués de presse est excessive, compte tenu du nombre de questions que nous possédons du corpus BCE-4. Notre participation au développement du langage TimeML et de ses outils d'annotation nous a permis d'envisager des façons d'exploiter l'information différentes de celles traditionnellement utilisées en question-réponse. Les méthodes de question-réponse traditionnelle basées sur la recherche et l'extraction d'information sont appropriées pour les questions factuelles ou dans notre problème comme une aide aux personnes offrant le service. Pour automatiser l'identification de la réponse, nous devons compter sur un système nous permettant d'interpréter le monde et d'y effectuer des raisonnements simples, comme nous le faisons avec le temps où nous ordonnons les événements selon différents aspects.

Chapitre 9

Conclusion

9.1 Réalisations

Dans la thèse nous avons décrit une architecture question-réponse pour automatiser la réponse aux courriels dans le cadre du service de relation avec les investisseurs. Nous avons solutionné certains problèmes pour réaliser la tâche, mais nous avons aussi soulevé de nouvelles questions. Pour terminer, nous analysons le rôle de notre architecture pour la mise en oeuvre de la réponse automatisée aux courriels et nous abordons quelques pistes de solutions pour améliorer les composantes de l'architecture.

Notre architecture question-réponse touche à toutes les étapes de la réponse automatisée aux courriels. Dans notre cas, nous avons surtout considéré le traitement du courriel, c'est-à-dire l'identification du besoin d'information de l'utilisateur et des moyens pour le satisfaire. Nous justifions l'attention que nous avons portée à cette tâche par le fait de donner une bonne réponse à une mauvaise interprétation du courriel n'est pas plus satisfaisant que de donner une mauvaise réponse à une bonne interprétation de celui-ci.

L'assemblage des composantes de traitement que nous avons étudiées permet de créer un système pour traiter le courriel, de la réception du courriel à la recherche de la réponse. Nous sommes en mesure d'extraire la question et l'information pertinente à la recherche de la réponse à partir du courriel, pour ensuite, à partir de la question et du contexte, déterminer le type d'information recherchée et le traitement à réaliser

pour trouver la réponse. Un tel système solutionnerait partiellement le problème de la réponse automatisée aux courriels, il nous resterait à identifier précisément la réponse à partir de ressources similaires à celles que nous avons décrites et à rédiger le courriel de réponse. Pour la rédaction de la réponse, nous pourrions tirer profit de la méthode exploitant le raisonnement à base de cas élaborée par Luc Lamontagne dans le cadre du projet Merkure [40]. Même sans ces deux étapes, le système pourrait améliorer la productivité des préposés assignés aux services d'information en diminuant le temps de traitement d'une requête.

Notre problème se situait initialement dans le cadre d'une collaboration entre BCE et notre laboratoire. Pour diverses raisons, il n'a malheureusement pas été possible de compléter la mise en oeuvre et l'évaluation à l'intérieur de ce service. Au moment d'évaluer le système, notre collaboration avec l'entreprise BCE était terminée et nos collaborateurs n'occupaient plus les mêmes fonctions. Nous n'avons donc pas pu obtenir les ressources nécessaires pour terminer l'implémentation du système et réaliser une évaluation significative de celui-ci. En plus des ressources mentionnées précédemment, l'évaluation du système aurait nécessité un corpus de courriels différent de celui que nous avons étudié et l'accès à des préposés prêts à évaluer la pertinence de l'analyse des courriels dans le cadre de leur tâche. L'évaluation d'une telle approche étant très subjective, nous n'avons pas de mesure pour déterminer le score d'un système de réponse automatisée comme c'est le cas pour les systèmes de question-réponse évalués lors de la conférence TREC. La difficulté d'évaluer notre travail réside dans le fait que nous travaillons avec des courriels diversifiés provenant d'utilisateurs, dans un cadre de traitement manuel, et non pas avec des requêtes modifiées pour éliminer certaines difficultés de traitement. Nous avons tout de même évalué séparément chaque composante de notre architecture.

Cette architecture question-réponse serait adaptable aux services d'information en général, en utilisant les méthodes de traitement et d'analyse présentées dans la thèse. Nos méthodes dépendent peu de la spécificité du domaine d'application, si ce n'est le cadre du service de relations avec des usagers. Ce cadre général nous permet ainsi d'envisager la mise en oeuvre de systèmes de réponse automatisée aux courriels avec l'architecture question-réponse en minimisant l'étape d'ingénierie cognitive. Sur ce point, notre approche se distingue des approches existantes parce que les approches générales ne font qu'un traitement simple du courriel et que les approches spécialisées nécessitent une connaissance exhaustive du domaine d'application.

Nous avons montré, au chapitre 4, que le repérage des questions d'un courriel dans un cadre spécialisé peut être réalisé par une méthode à base de règles. Elle est adaptée au traitement des questions provenant d'utilisateurs, comme celui des services de référence, et elle n'intègre pas de connaissances particulières des domaines de la finance ou de l'investissement. L'intégration de cette méthode polyvalente de repérage des questions avec les autres étapes de traitement permet de déterminer une partie de l'information recherchée par un utilisateur. Notre méthode d'analyse du courriel est différente de ce qui est fait dans le repérage des pourriels ou du raisonnement à base de cas parce nous n'utilisons pas de données spécifiques au problème traité. Cette indépendance entre les données du problème et son traitement nous permet d'utiliser l'architecture de la solution indépendamment de la spécialisation du domaine d'application.

Lors de l'étude de la question (par sa classification, sa modélisation et sa description) nous avons établi des liens entre : les sources de réponses, les taxonomies de questions et les niveaux d'analyse linguistique. Ces relations sont pertinentes parce que les questions peuvent être classifiées selon plusieurs critères. Le type des sources utilisées pour répondre au courriel est déterminé par les classes dans lesquelles nous classifions la question. La qualité des réponses suggérées dépend essentiellement du niveau d'analyse linguistique que nous sommes en mesure d'y appliquer. En augmentant le niveau d'analyse linguistique de la question, nous classifions la question en fonction d'un plus grand nombre de taxonomies et nous pouvons en extraire une formalisation plus riche qui tient compte d'une information plus complète.

Dans la réponse automatisée aux courriels pour les services de relations avec les investisseurs, nous devons exploiter des données textuelles et des données structurées pour trouver la réponse aux questions. Nous avons exploité le repérage des rôles sémantiques pour analyser les questions et les documents candidats, dans le but d'extraire les relations entre les entités induites par les verbes. Le repérage des rôles sémantiques d'une question permet d'identifier les rôles manquants qui correspondent à l'information demandée par un utilisateur. Cette représentation de la question est ensuite utilisée pour récupérer la réponse dans les données structurées ou les documents pertinents, grâce à un appariement des entités avec les relations déjà connues.

Nous avons aussi contribué au développement du langage d'annotation des expressions temporelles et des événements TimeML. Dans un premier temps, nous avons établi qu'une représentation graphique des événements et de leurs relations facilite l'annotation d'un document avec TimeML. La représentation graphique permet à l'annotateur d'identifier les relations entre les événements dans un contexte plus large parce que la représentation des événements lui donne un point de vue global du récit. Cette idée est essentielle au développement de l'outil d'annotation semi-automatique TANGO. Avec TANGO, l'annotateur peut manipuler les marqueurs temporels et les événements, directement à partir de la représentation graphique, pour annoter les relations.

Le système d'information que nous proposons pour répondre automatiquement aux courriels permettra d'améliorer la qualité des services avec les usagers en suggérant aux préposés des façons d'analyser le courriel. En ajoutant une étape pour récupérer la réponse avec certitude et une étape pour générer le texte d'une réponse, nous pourrons construire un système complet de réponse automatisée aux courriels. Mais, les systèmes de réponse automatisée aux courriels ne pourront pas remplacer le traitement manuel des courriels tant que les méthodes de traitement de la langue naturelle ne performeront pas au même niveau qu'un expert.

9.2 Discussion et travaux futurs

Nous avons réalisé les étapes de traitement et d'analyse dans un cadre expérimental, mais dans le but d'implémenter un système complet de réponse automatisée aux courriels. Nous pouvons tenter d'améliorer chaque étape de traitement en solutionnant les problèmes que nous avons rencontrés par des méthodes ou des points de vue différents.

Puisque nous travaillons avec des courriels provenant d'investisseurs au profil diversifié, les questions contenues dans les courriels sont souvent ambiguës, un problème que nous n'avons pas traité spécifiquement dans notre travail. Une manière de désambiguïser les questions est d'utiliser l'approche de la question-réponse coopérative [8]. Cette approche nous servirait à restreindre l'information que l'utilisateur peut demander à celle mise à notre disposition. La question-réponse coopérative permet de gérer les situations où la réponse ne se trouve pas dans les données disponibles. Ce

type de système retournera une réponse utile pour expliquer à l'utilisateur quels éléments de sa requête font en sorte qu'elle ne peut pas être répondue. Pour les préposés et les utilisateurs, il est plus pratique d'identifier les incongruités et les ambiguïtés de la requête que de retourner un ensemble d'informations erronées.

Dans le chapitre 5, nous avons exploré des approches pour comprendre les intentions sous-jacentes à la formulation d'une question. Nous avons étudié des taxonomies de questions qui correspondent à des niveaux d'analyse linguistique. Dans ce travail, nous n'avons réalisé des expériences que sur les aspects syntaxique et sémantique de l'analyse linguistique de la question. Mais pour comprendre une question, il faut l'analyser en tenant compte du discours et des règles de communication implicites au contexte habituellement associées à l'analyse pragmatique du message. Le développement de méthodes d'analyse automatique pour ces niveaux linguistiques rendrait l'analyse de la question plus efficace. Les taxonomies de question, en relation avec les niveaux d'analyse linguistique, pourraient être étudiées pour évaluer la complexité d'une question en fonction de la classe à laquelle elle appartient.

Une autre piste de solution pour améliorer la compréhension de la question serait d'améliorer les formalismes de représentation de questions pour les utiliser dans des modèles algorithmiques. La complexité et le niveau d'abstraction associés aux modèles formels de la question constituent le principal obstacle à leur utilisation pour l'analyse des courriels.

Les méthodes statistiques que nous avons utilisées sont dépendantes des données sur lesquelles les statistiques sont extraites. Avec des ressources appropriées pour nos tâches, nos résultats seraient meilleurs. Pour le traitement des questions, le seul corpus accessible est la collection de questions de la conférence TREC ; pour le traitement des courriels, nous n'avons que le corpus ENRON ; mais ces deux corpus ne contiennent pas de courriels de nature interrogative. Pour améliorer l'analyse des courriels, nous pourrions créer un corpus de courriels contenant des questions annotées avec de l'information syntaxique et sémantique nécessaire à l'extraction d'attributs nous permettant d'utiliser des méthodes statistiques. En exploitant de nouveaux attributs linguistiques plus discriminants, nous pourrons améliorer les résultats obtenus par les méthodes statistiques.

Dans cette thèse, nous avons utilisé les machines à vecteur de support (SVM) pour réaliser des expériences basées sur des méthodes statistiques. L'algorithme des

SVM produit de très bons résultats lorsque les données d'entraînement forment deux classes équilibrées. Par contre lorsqu'une classe contient beaucoup plus d'exemples que l'autre, les résultats obtenus par l'algorithme sont décevants. Dans notre tâche de repérage des rôles sémantiques, plus de 90% des constituants n'ont pas de rôles sémantiques. Pour pallier à ce problème, nous devons exploiter des méthodes statistiques d'apprentissage qui soient stables pour les ensembles de données déséquilibrées. Nous pourrons aussi améliorer le repérage en combinant l'approche statistique avec un traitement à base de règles permettant de sélectionner ou d'éliminer des constituants en fonction des connaissances linguistiques.

Nous ne traitons pas en profondeur l'extraction de la réponse dans notre problème parce que nous ne disposons pas de données suffisantes pour l'aborder. Nous avons tenté, par le développement d'une ontologie, de modéliser les connaissances du domaine, mais ce n'est pas suffisant. Pour répondre correctement aux courriels, nous devons disposer de connaissances touchant plusieurs niveaux conceptuels et être en mesure de les exploiter en fonction des tâches à réaliser. Une des motivations derrière le développement de l'ontologie du domaine était d'exploiter la popularité du formalisme OWL en transformant une question dans le langage de requête OWL-QL qui aurait été utilisé pour interroger l'ontologie. Cette étape pourrait être réalisée dans le cadre de travaux futurs à condition de pouvoir formaliser précisément une question et les données du domaine.

Le problème de la réponse automatisée aux courriels pour les services d'information a un aspect *artificiel*. Les techniques de question-réponse factuelle nous permettent de répondre aux questions simples à partir d'une grande collection de ressources textuelles. Un problème que nous avons rencontré avec les courriels du service de relation avec les investisseurs, vient du fait que l'information recherchée n'est pas accessible à l'utilisateur. Si cette information était accessible, l'utilisateur pourrait combler son besoin d'information par une recherche dans la base documentaire, comme le site web de l'entreprise. Puisque l'information n'est pas accessible, le service d'information se retrouve à répondre à des courriels concernant des informations factuelles que l'utilisateur pourrait lui-même récupérer. Pour désengorger les services d'information, nous pouvons rendre l'information plus accessible : en améliorant la structure et la présentation de l'information, en utilisant des repères sémantiques lors de la création de l'information et en fournissant à l'utilisateur des outils de recherche

d'information ; ou nous utilisons un système de réponse automatisée aux courriels. Ces solutions permettent de dégager des ressources qui répondront aux questions plus complexes pour lesquelles nous n'avons pas d'outil présentement pour y répondre automatiquement.

9.3 Conclusion

L'approche question-réponse que nous avons exploiter dans nos travaux pour traiter le problème de la réponse automatisée aux courriels est pertinente pour l'automatisation des services d'information, particulièrement dans la gestion des relations avec les usagers. La combinaison des étapes de traitement présentées dans cette thèse constituent une architecture pour construire un système de réponse automatisée aux courriels. L'implémentation d'un tel système d'information nécessite la résolution de problèmes de natures appliquée et théorique. Les expériences que nous avons menées constituent un point de départ pour la résolution des problèmes pratiques. L'étude des aspects théoriques de la question est essentielle à la mise en oeuvre d'un tel système, parce que les façons de concevoir les problèmes influencent les approches utilisées pour développer les étapes de traitement.

L'architecture question-réponse pour répondre automatiquement aux courriels est une suite d'étapes de traitement interdépendantes. Les expériences réalisées dans ma thèse concernent principalement les méthodes d'analyse des courriels. Si un système de réponse automatisée aux courriels basé sur la question-réponse n'identifie pas correctement ce qui est demandé par l'utilisateur, alors il est impossible de retourner une information pertinente dans la réponse. Même si nous n'avons pas implémenté et évalué une méthode pour générer une réponse, nous pouvons utiliser l'analyse des courriels comme un système d'aide à la réponse, qui limite les domaines où chercher la réponse et suggère des manières d'y répondre. La réponse automatisée aux courriels reste un problème difficile à résoudre nécessitant la compréhension du langage et une capacité de raisonnement similaire à celle d'un expert du domaine d'application.

Bibliographie

[1] AKBANI, Rehan, Stephen KWEK et Nathalie JAPKOWICZ. « Applying support vector machines to imbalanced datasets ». Dans *Proceedings of the 15th european conference on machine learning (ECML)*, 2004, p. 39–50.

[2] AONE, Chinatsu et Mila RAMOS-SANTACRUZ. « REES : A large-scale relation and event extraction system ». Dans *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, 2000.

[3] AUSTIN, John Langshaw. *How to do things with words.* Oxford University Press, Oxford, 1962.

[4] BAKER, Collins F., Charles J. FILLMORE et John B. LOWE. « The Berkeley FrameNet project ». Dans *Proceedings of the COLING-ACL.* Montreal, Canada, 1998.

[5] BANTER INC. *Natural Language Engines for Advanced Customer Interaction.* Rapport technique, Banter Inc., 2001.

[6] BÉLANGER, Luc et Guy LAPALME. « Identification de questions pour traiter les courriels par une méthode question-réponse ». Dans PRUNELLE, Gerald, Cédrick FAIRON et Anne DISTER (rédacteurs), *Le poids des mots : Actes des 7es journées internationales d'analyse statistique des données textuelles*, 2004, p. 128–135.

[7] BÉLANGER, Luc et Guy LAPALME. « Identification des rôles sémantiques par la classification ». Dans MAKARENKOV, Vladimir, Guy CUCUMEL et Francois-Joseph LAPOINTE (rédacteurs), *Comptes rendus des 12es Rencontres de la Société Francophone de Classification.* Société Francophone de Classification, Montréal, Canada, mai 2005, p. 59–62.

[8] BENAMARA, Farah et Patrick SAINT DIZIER. « Advanced relaxation for cooperative question answering ». Dans MAYBURY, Mark T. (rédacteur), *New Directions in Question Answering*, chapitre 21. AAAI Press, 2004, p. 234–254.

[9] BOGURAEV, Branimir, Jose CASTAÑO, Rob GAIZAUSKAS, Bob INGRIA, Graham KATZ, Bob KNIPPEN, Jessica LITTMAN, Inderjeet MANI, James PUSTEJOVSKY, Antonio SANFILIPPO, Andrew SEE, Andrea SETZER, Roser SAURÍ, Amber STUBBS, Beth SUNDHEIM, Svetlana SYMONENKO et Marc VERHAGEN. *TimeML 1.2.1 : A Formal Specification Language for Events and Temporal Expressions*, 2005.

[10] BURKE, Robin D., Kristian J. HAMMOND, Vladimir KULYUKIN, Steven L. LYTINEN, Noriko TOMURO et Scott SCHOENBERG. « Question answering from frequently asked question files : Experiences with the FAQ FINDER system ». *AI Magazine*, tome 18, n° 2, juin 1997, p. 57–66.

[11] CARBONELL, Jaime, Eduard HOVY, Donna HARMAN, Steve MAIORANO, John PRANGE et Karen SPARCK-JONES. *Vision Statement to Guide Research in Question & Answering (Q&A) ans Text Summarization.* Rapport technique, NIST, 2000.

[12] CARRERAS, Xavier et Lluís MÀRQUEZ. « Introduction to the CoNLL-2004 shared task : Semantic role labeling ». Dans NG, Hwee Tou et Ellen RILOFF (rédacteurs), *HLT-NAACL 2004 Workshop : Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Association for Computational Linguistics, Boston, Massachusetts, USA, mai 2004, p. 89–97.

[13] CHANG, Chih-Chung et Chih-Jen LIN. *LIBSVM : a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[14] CHAWLA, Nitesh, Nathalie JAPKOWICZ et Alek KOLCZ. « Editorial : Special issue on learning from imbalanced data sets ». *ACM SIGKDD Explorations*, tome 6, n° 1, juin 2004, p. 1–6.

[15] CHEN, John et Owen RAMBOW. « Use of deep linguistic features for the recognition and labeling of semantic arguments ». Dans *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003.

[16] COLLINS, Michael. *Head-Driven Statistical Models for Natural Language Parsing.* Thèse de doctorat, University of Pennsylvania, Philadelphia, PA, 1999.

[17] CUNNINGHAM, Hamish, Diana MAYNARD, Kalina BONTCHEVA et Valentin TABLAN. « GATE : A framework and graphical development environment for robust NLP tools and applications ». Dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphie, Pennsylvanie, juillet 2002.

[18] DAY, David, John ABERDEEN, Lynette HIRSCHMAN, Robyn KOZIEROK, Patricia ROBINSON et Marc VILAIN. « Mixed-initiative development of language processing systems ». Dans *Fifth conference on applied natural language systems.* Washington D.C., 1997, p. 88–95.

[19] DUAN, Kai-Bo et S. Sathiya KEERTHI. « Which is the best multiclass SVM method ? an empirical study ». Dans OZA, Nikunj C., Robi POLIKAR, Josef KITTLER et Fabio ROLI (rédacteurs), *Multiple Classifier Systems : 6th International Workshop, MCS 2005, Seaside, CA, USA, June 13-15, 2005.*, *Lecture Notes in Computer Science*, tome 3541, 2005, p. 278–285.

[20] DUBOIS, Julien. *Classification automatique de courrier électronique.* Mémoire de maîtrise, Université de Montréal, juin 2002.

[21] GILDEA, Daniel et Daniel JURAFSKY. « Automatic labeling of semantic roles ». Dans *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00)*. Association for Computational Linguistics, Hong-Kong, octobre 2000, p. 512–520.

[22] GILDEA, Daniel et Daniel JURAFSKY. « Automatic labeling of semantic roles ». *Computational Linguistics*, tome 28, n° 3, 2002, p. 245–288.

[23] GILDEA, Daniel et Martha PALMER. « The necessity of syntactic parsing for predicate argument recognition ». Dans *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02)*. Philadelphia, PA, 2002, p. 308–314.

[24] GREEN, JR., Bert F., Alice K. WOLF, Carol CHOMSKY et Kenneth LAUGHERY. « Baseball : an automatic question answerer ». Dans *Computers & thought*. McGraw-Hill, 1963, p. 207–216.

[25] GROENENDIJK, Jeroen et Martin STOKHOF. « Questions ». Dans VAN BENTHEM, Johan et Alice TER MEULEN (rédacteurs). *Handbook of Logic & Language*, chapitre 19. MIT Press - North Holland, 1997, p. 1055–1124.

[26] HAMBLIN, Charles L. « Questions ». *Australasian Journal of Philosophy*, tome 36, 1958, p. 159–168.

[27] HARABAGIU, Sanda, Dan MOLDOVAN, M. PAŞCA, R. MIHALCEA, M. SURDEANU, R. GÎRJI, V. RUS et P. MORÂRESCU. « Falcon : Boosting knowledge for answer engines ». Dans *Proceedings of the ninth text retrieval conference (TREC-9)*. NIST, Gaithersburg, Md, 2001.

[28] HERMJAKOB, Ulf. « Parsing and question classification for question answering ». Dans *Proceedings of the Association for Computational Linguistics 2001 Workshop on Open-Domain Question Answering*. Association for Computational Linguistics, 2001, p. 17–22.

[29] HOVY, Eduard H., Ulf HERMJAKOB et Chin-Yew LIN. « The use of external knowledge of factoid QA ». Dans *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[30] HSU, Chih-Wei, Chih-Chung CHANG et Chih-Jen LIN. *A practical guide to support vector classification*. Rapport technique, Department of Computer Science and Information Engineering, Natioanl Taiwan University, Taipei 106, Taiwan, 2003.

[31] JOACHIMS, Thorsten. « Making large-scale SVM learning practical ». Dans SCHÖLKOPF, Bernhard, Chris BURGES et Alex J. SMOLA (rédacteurs), *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, Cambridge, MA, 1999, p. 41–56.

[32] JUPITER COMMUNICATIONS. *E-mail Customer Service : Taking Control of Rising Customer Demand*. Rapport technique, Jupiter Communications, 2000.

[33] KATZ, Boris, Sue FELSHIN, Jimmy LIN et Gregory MARTON. « Viewing the web as a virtual database ». Dans MAYBURY, Mark T. (rédacteur), *New Directions In Question Answering*, chapitre 16. MIT Press, Cambridge, Massachussets, 2004, p. 215–226.

[34] KATZ, Boris, Sue FELSHIN, Deniz YURET, Ali IBRAHIM, Jimmy LIN, Gregory MARTON, Alton Jerome MCFARLAND et Baris TEMELKURAN. « Omnibase : Uniform access to heterogeneous data for question answering ». Dans *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002)*. Stockholm, Sweden, juin 2002.

[35] KINGSBURY, Paul et Martha PALMER. « From Treebank to PropBank ». Dans *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*. Las Palmas, Spain, 2002.

[36] KLIMT, Bryan et Yiming YANG. « The enron corpus : A new dataset for email classification research ». Dans *First Conference on Email and Anti-Spam (CEAS)*. Mountain View, CA, juillet 2004.

[37] KOSSEIM, Leila et Guy LAPALME. « Critères de sélection d'une approche pour le suivi automatique du courriel ». Dans *Actes de la 8e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*. Tours, France, juillet 2001, p. 357–371.

[38] KOSSEIM, Leila, Luc PLAMONDON et Guy LAPALME. « La réponse automatique comme solution à la gestion des relations avec la clientèle ». *Revues des sciences et technologies de l'information (RSTI) série Ingénierie des systèmes d'information (ISI)*, tome 8, n° 3, 2003, p. 91–114.

[39] KWOK, Kui-Lam, Laszlo GRUNFELD, Norbert DINSTL et M. CHAN. « Trec2001 question-answerm web and cross language experiments using pircs ». Dans *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[40] LAMONTAGNE, Luc. *Une approche CBR textuel de réponse au courrier électronique*. Thèse de doctorat, Université de Montréal, juillet 2004.

[41] LENHERT, Wendy. *The Process of question answering : a computer simulation of cognition*. Erlbaum Associates, New York, 1978.

[42] MANI, Inderjeet et Goeorges WILSON. « Robust temporal processing of news ». Dans *Proceedings of the ACL'2000 Conference*. Hong Kong, octobre 2000.

[43] MAYBURY, Mark T., David DAY, John PRANGE, James PUSTEJOVSKY et Janice WIEBE (rédacteurs). *New Directions in Question Anwering : Papers from the 2003 AAAI Spring Symposium*. American Association for Artificial Intelligence, Menlo Park, Calif., 2003.

[44] MEYERS, Adam, Ruth REEVES, Catherine MACLEOD, Rachel SZEKELY, Veronika ZIELINSKA, Brian YOUNG et Ralph GRISHMAN. « The nombank project : An interim report ». Dans MEYERS, Adam (rédacteur), *HLT-NAACL 2004*

Workshop : Frontiers in Corpus Annotation. Association for Computational Linguistics, Boston, Massachusetts, USA, mai 2004, p. 24–31.

[45] MOLDOVAN, Dan, Sanda HARABAGIU, M. PAŞCA, R. MIHALCEA, R. GOODRUM, R. GÎRJI et V. RUS. « Lasso : A tool for surfing the answer net ». Dans *Proceedings of the eight text retrieval conference (TREC-8)*. NIST, 2000.

[46] MOSCHITTI, Alessandro et Cosmin Adrian BEJAN. « A semantic kernel for predicate argument classification ». Dans NG, Hwee Tou et Ellen RILOFF (rédacteurs), *HLT-NAACL 2004 Workshop : Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Association for Computational Linguistics, Boston, Massachusetts, USA, mai 2004, p. 17–24.

[47] NARAYANAN, Srini et Sanda HARABAGIU. « Answering questions using advanced semantics and probabilistic inference ». Dans *Proceedings of the Workshop on pragmatics of question-answering at HTL-NAACL 2004*. Association for Computational Linguistics, 2004, p. 10–16.

[48] PALMER, Martha, Daniel GILDEA et Paul KINGSBURY. « The proposition bank : An annotated corpus of semantic roles ». *Computational Linguistics*, tome 31, n° 1, 2005, p. 71–106.

[49] PALMER, Martha, Joseph ROSENZWEIG et Scott COTTON. « Automatic predicate argument analysis of the Penn TreeBank ». Dans ALLAN, James (rédacteur), *Proceedings of the First International Conference on Human Language Technology Research*. Morgan Kaufmann, San Francisco, CA, 2001.

[50] PLAMONDON, Luc. *Le système de question-réponse QUANTUM*. Mémoire de maîtrise, Université de Montréal, mars 2002.

[51] PLAMONDON, Luc, Guy LAPALME et Leila KOSSEIM. « The QUANTUM question answering system at trec 11 ». Dans VOORHEES, Ellen M. et Lori P. BUCKLAND (rédacteurs), *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*. NIST, 2002.

[52] POMERANTZ, Jeffrey. *Question Taxonomies for Digital Reference*. Thèse de doctorat, Syracuse University, juillet 2003.

[53] PRADHAN, Samer, Kadri HACIOGLU, Wayne WARD, James H. MARTIN et Daniel JURAFSKY. « Support vector learning for semantic argument classification ». *Machine Learning*, tome 60, n° 1, 2005, p. 11–39.

[54] PRADHAN, Samer, Wayne WARD, Kadri HACIOGLU, James H. MARTIN et Dan JURAFSKY. « Shallow semantic parsing using support vector machines ». Dans *Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistic annual meeting*. Association for Computational Linguistics, Boston, MA, mai 2004.

[55] PRIOR, Mary et Arthur PRIOR. « Erotetic logic ». *Philosophical Review*, tome 64, n° 1, 1955, p. 43–59.

[56] PUNYAKANOK, Vasin, Dan ROTH et Wen-Tau YIH. « The necessity of syntactic parsing for semantic role labeling ». Dans KAELBLING, Leslie Pack et Alessandro SAFFIOTTI (rédacteurs), *IJCAI*. Professional Book Center, 2005. ISBN 0938075934, p. 1117–1123.

[57] PUSTEJOVSKY, James, Luc BÉLANGER, José CASTAÑO, Rob GAIZAUSKAS, Patrick HANKS, Bob INGRIA, Graham KATZ, Dragomir RADEV, Anna RUMSHISKY, Antonio SANFILIPPO, Roser SAURÍ, Andrea SETZER, Beth SUNDHEIM et Marc VERHAGEN. *NRRC Summer Workshop on Temporal and Event Recognition for Question Answering Systems : Final Report.* Rapport technique, NRRC, MITRE Corporation, Bedford, Mass., 2002.

[58] PUSTEJOVSKY, James, Inderjeet MANI, Luc BÉLANGER, Branimir BOGURAEV, Bob KNIPPEN, Jessica LITTMAN, Anna RUMSHISKY, Andrew SEE, Svetlana SYMONEN, James Van GUILDER, Linda Van GUILDER et Marc VERHAGEN. *ARDA Summer Workshop on Graphical Annotation Toolkit for TimeML.* Rapport technique, NNRC, MITRE Corporation, Bedford, Mass., 2003.

[59] RIFKIN, Ryan et Aldebaro KLAUTAU. « In defense of one-vs-all classification ». *Journal of Machine Learning Research*, tome 5, 2004, p. 101–141.

[60] RUPPENHOFER, Josef, Michael ELLSWORTH, Miriam R. L. PETRUCK et Christopher R. JOHNSON. *FrameNet : Theory and Practice.* Berkeley FrameNet project, juin 2005.

[61] SEARLE, John. *Speech acts.* Cambridge University Press, Cambridge, 1969.

[62] SHORTLIFFE, Edward H. *Computer-Based Medical Consultations : MYCIN.* Elsevier/North-Holland, New York, 1976.

[63] SURDEANU, Mihai, Sanda HARABAGIU, John WILLIAMS et Paul AARSETH. « Using predicate-arguments structures for information extraction ». Dans *Proceedings of ACL-2003.* Association for Computational Linguistics, 2003, p. 8–15.

[64] THE RADICATI GROUP, INC. *Taming the growth of email : An ROI analysis.* White Paper, mai 2005.

[65] TURING, Alan M. « Computing machinery and intelligence ». Dans FEIGENBAUM, Edward A. et Julien FELDMAN (rédacteurs), *Computers & thought.* McGraw-Hill, 1963, p. 11–38.

[66] VERHAGEN, Marc. *Times Between The Lines.* Thèse de doctorat, Brandeis University, novembre 2004.

[67] VOORHEES, Ellen M. « Question answerin in TREC ». Dans Voorhees et Harman [69], chapitre 10, p. 233–257.

[68] VOORHEES, Ellen M. et Lori P. BUCKLAND (rédacteurs). *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004).* NIST, Department of Commerce, National Institute of Standards and Technology, 2004.

[69] VOORHEES, Ellen M. et Donna K. HARMAN (rédacteurs). *TREC : Experiment and evaluation in information retrieval*. MIT Press, 2005.

[70] WATANABE, Yasuhiko, Kazuya YOKOMIZO et Yoshihiro OKADA. « A question answer system using mail posted to a mailing list ». Dans KEŠELJ, Vlado et Tsutomu ENDO (rédacteurs), *Proceedings of the PACLING'03*, 2003, p. 335–342.

[71] WOODS, Williams A., Richard M. KAPLAN et Bonnie Lynn NASH-WEBBER. *The Lunar Sciences Natural Language Information System : Final Report*. BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, MA, juin 1972.

[72] XUE, Nianwen et Martha PALMER. « Calibrating features for semantic role labeling ». Dans *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

Annexe A

TimeML

A.1 Le langage TimeML

Le langage d'annotation TimeML a été développé pour étiqueter l'information temporelle dans les textes dans le but de répondre aux questions concernant les événements et les entités temporelles. L'annotation de l'information temporelle directement dans le texte avec TimeML permet d'identifier les représentations temporelles et les compléments d'information pouvant déjà être annotés dans le texte. Nous avons participé au développement de TimeML dans le but de solutionner quatre problèmes liés à l'identification des événements et de la temporalité :

1. l'ancrage absolu des événements dans le temps ;
2. l'ordonnancement chronologique et relatif des événements entre eux ;
3. le raisonnement à partir d'expressions temporelles sous-spécifiées dans le contexte d'énonciation ;
4. le raisonnement sur la persistance des événements.

Les premières versions du langage TimeML ont été développées lors de deux projets de recherche réalisés sous la forme d'un atelier de travail (*workshop*) : TERQAS [57] et TANGO [58], auxquels nous avons participé. Notre principale contribution à ces travaux est un outil pour générer une représentation graphique d'un document TimeML, qui à été développé lors de TERQAS et que nous avons utilisé comme prototype lors du développement de l'outil d'annotation TANGO. La création d'un langage d'an-

notation et d'un outil d'annotation a permis la création d'un corpus démontrant le bien-fondé de l'annotation de documents avec TimeML. Notre contribution à la création du corpus, lors de la tâche d'annotation, s'est manifestée sous la forme d'outils pour aider les annotateurs, puisqu'aucun logiciel d'annotation n'était adapté pour cette tâche. Le corpus TimeBank se présente en deux versions.

1. **TimeBank 1.1** contient 186 articles de nouvelles annotés avec la version 1.1 du langage TimeML lors de son développement. Ce corpus a été assemblé par plusieurs annotateurs et des documents ont dû être exclus de cette version parce que l'annotation n'était pas utilisable. Dans ce corpus, il y a des documents qui doivent être corrigés, car on y retrouve des erreurs et des omissions. Dans ce document, nous faisons référence à cette version du corpus.
2. **AQUAINT TimeML** contient 81 articles de nouvelles annotés par un des trois annotateurs de Brandeis University ou de Georgetown University. Tous les articles de ce corpus correspondent à la version 1.2.1 de TimeML.

Le langage TimeML est construit autour des concepts d'objet et de relation temporels. Les objets temporels sont les événements et les expressions temporelles explicites. Les relations temporelles sont identifiées par des *liens* et divisées en trois classes : liens temporels, liens subordonnants, liens aspectuels.

L'annotation des objets temporels se fait directement dans le texte par des balises XML autour de l'entité textuelle décrivant l'objet. Les deux types d'objets temporels principaux sont :

- `<EVENT>` pour annoter les événements ;
- `<TIMEX3>` pour annoter les expressions temporelles explicites, déictiques et relatives, p. ex. *juin 2005, aujourd'hui, dimanche passé.*

Un autre type d'objet est aussi considéré pour annoter des sections de textes indiquant comment les objets temporels sont liés entre eux : le signal `<SIGNAL>`. Les objets étiquetés `<EVENT>` sont généralement des verbes, mais un nom peut aussi être utilisé pour dénoter un événement, comme un *attentat* pour dénoter l'événement *des personnes sont mortes dans l'attentat du 11 septembre 2001.* Dans certains cas, un événement est aussi exprimé par un adjectif, une clause prédicative ou une préposition, mais les réalisations des événements ainsi exprimés sont plus difficiles à annoter. L'étiquette `<EVENT>` est aussi utilisée dans certains cas pour identifier des états de transition lorsqu'ils ont un lien dans le récit. Les événements étiquetés sont divisés en sept

classes :

occurence : la plupart des événements font partie de cette classe, ils décrivent ce qui se produit dans le monde ;

state : les états décrivant les circonstances dans lequel un événement a lieu et dont l'état peut être modifié ; et les états introduits par les *i-action*, *i-state* et *reporting* ;

reporting : description de l'action d'une personne par un acte narratif ;

i-action : une action intentionnelle introduisant un autre événement, comme un essai, une enquête, un report, un ordre, une demande, une promesse, une nomination ;

i-state : similaire à *i-action* mais pour identifier un état tel que penser, ressentir, suspecter, douter, vouloir, désirer, détester, être prêt, être capable ;

aspectual : un événement débutant, terminant ou continuant une action ;

perception : constatation physique d'un événement telle qu'entendre ou voir l'action.

TimeML fait une distinction entre l'événement et sa réalisation, qui se caractérise par le temps et l'aspect. La distinction entre l'événement et sa réalisation est nécessaire lorsqu'un événement est réalisé plus d'une fois, comme dans le cas d'un événement périodique.

Les liens temporels sont définis entre les événements et les expressions temporelles. Dans l'annotation, les `<EVENT>` ne participe jamais à une relation, c'est la réalisation (`<MAKEINSTANCE>`) de l'événement qui y participe.

Les **liens temporels** `<TLINK>` représentent la relation entre deux objets temporels, que ce soit deux événements, deux marqueurs temporels ou un marqueur temporel et un événement. Il y a quatorze types de relations identifiées par les `<TLINK>`, bien que certaines soient simplement l'inverse d'une autre :

before et after spécifient qu'un objet temporel précède ou suit l'autre objet temporel de la relation ;

ibefore et iafter spécifient qu'un objet temporel est immédiatement avant ou après un autre ;

includes et is-included spécifient qu'un objet temporel inclut ou est inclus dans un autre, p. ex. *John* ***arrived*** *in Montreal* ***yesterday***. ;

during spécifie que l'état ou l'événement se poursuit durant une période de temps, p. ex. *John **taught** for **90 minutes***;

during-inv est l'inverse de la relation précédente ;

simultaneous spécifie que deux instances d'événement semblent coïncider dans le temps ;

identity indique que deux objets temporels représentent le même événement ;

begins spécifie qu'un événement débute par l'objet temporel avec lequel il est lié ;

begun-by est l'inverse de **begin**, elle relie un objet temporel à un événement débutant par l'objet temporel ;

ends et ended-by sont similaires aux deux relations précédentes sauf qu'elles spécifient la fin de l'événement.

Les **liens subordonnants** `<SLINK>` identifient les relations entre deux événements ou entre un événement et un signal. Ils sont habituellement introduits par des verbes modaux qui impliquent une confirmation, une validation ou une invalidation par exemple. Les liens subordonnants sont définis selon six types de relations qui interagissent avec les classes d'événements *reporting*, *i-state* et *i-action* :

modal introduit la possibilité d'un événement, p. ex. *John **promised** Mary to **buy** some beer*

evidential introduit la perception ou le compte-rendu de l'événement, p. ex. *John **said** he **bought** a pack of beer.*

neg-evidential introduit la perception ou rapporte que l'événement ne s'est pas réalisé, p. ex. *John **denied** he **bought** beers*

factive est une action qui implique ou présuppose qu'un événement a déjà eu lieu, p. ex. *John **managed** to **leave** the party.*

counter-factive est la négation de la relation précédente p. ex. *John **forgot** to **buy** beers.*

conditional indique que la réalisation de l'action entraînera l'événement en relation.

Les **liens aspectuels** `<ALINK>` mettent en relation un aspect de la relation qui existe entre deux événements. Les liens aspectuels sont un hybride des liens temporels et subordonnants, car ils caractérisent simultanément une relation temporelle et une subordination aspectuelle. Les cinq types de relations identifiées sont :

initiates p. ex. *John **started** to **read** his newspaper.*

culminates p. ex. *John **finished reading** his newspaper.*

terminates p. ex. *John **stopped reading** his newspaper.*

continue p. ex. *John **kept reading**.*

reinitiates p. ex. *John **returned reading** his newspaper.*

L'annotation avec TimeML de l'extrait présenté précédemment (p.111), nous donne l'annotation suivante :

```
<xml>
  <s>
    ISTANBUL, Turkey (AP) - Some 1,500 ethnic Albanians
    <EVENT eid="e2" class="OCCURENCE"> marched </EVENT>
    <TIMEX3 tid="t68" type="DATE" temporalFunction="true" functionInDocument="NONE"
      value="1998-03-08"> Sunday </TIMEX3>
    in downtown Istanbul,
    <EVENT eid="e62" class="I_ACTION"> burning </EVENT>
    Serbian flags
    <SIGNAL sid="s4"> to </SIGNAL>
    <EVENT eid="e47" class="I_STATE"> protest </EVENT>
    the
    <EVENT eid="e75" class="OCCURRENCE"> killings </EVENT>
    of ethnic Albanians by Serb police in southern Serb Kosovo province.
  </s>
  <s>
    The police
    <EVENT eid="e76" class="I_ACTION"> barred </EVENT>
    the crowd
    <SIGNAL sid="s9"> from </SIGNAL>
    <EVENT eid="e49" class="OCCURRENCE"> reaching </EVENT>
    the Yugoslavian consulate in downtown Istanbul, but
    <EVENT eid="e77" class="I_ACTION"> allowed </EVENT>
    them
    <SIGNAL sid="s12"> to </SIGNAL>
    <EVENT eid="e13" class="STATE"> demonstrate </EVENT>
    on nearby streets.
  </s>
  <MAKEINSTANCE aspect="NONE" eiid="ei193" tense="PAST" eventID="e2" />
  <MAKEINSTANCE aspect="NONE" eiid="ei194" tense="NONE" eventID="e62" />
  <MAKEINSTANCE aspect="NONE" eiid="ei195" tense="NONE" eventID="e47" />
  <MAKEINSTANCE aspect="PROGRESSIVE" eiid="ei196" tense="NONE" eventID="e75" />
```

```
  <MAKEINSTANCE aspect="NONE" eiid="ei197" tense="PAST" eventID="e76" />
  <MAKEINSTANCE aspect="PROGRESSIVE" eiid="ei198" tense="NONE" eventID="e49" />
  <MAKEINSTANCE aspect="NONE" eiid="ei199" tense="PAST" eventID="e77" />
  <MAKEINSTANCE aspect="NONE" eiid="ei200" tense="NONE" eventID="e13" />
  <TLINK relatedToTime="t68" eventInstanceID="ei193" relType="IS_INCLUDED" />
  <TLINK relatedToEventInstance="ei194" eventInstanceID="ei193"
    relType="SIMULTANEOUS" />
  <TLINK relatedToEventInstance="ei199" eventInstanceID="ei197"
    relType="SIMULTANEOUS" />
  <SLINK signalID="s4" subordinatedEventInstance="ei195" eventInstanceID="ei194"
    relType="FACTIVE"/>
  <SLINK subordinatedEventInstance="ei196" eventInstanceID="ei195" relType="MODAL" />
  <SLINK signalID="s9" subordinatedEventInstance="ei198" eventInstanceID="ei197"
    relType="COUNTER_FACTIVE"/>
  <SLINK signalID="s12" subordinatedEventInstance="ei200" eventInstanceID="ei199"
    relType="FACTIVE"/>
</xml>
```

L'information identifiée du texte est essentielle pour traiter l'information temporelle dans les systèmes de question-réponse et, aussi, pour la génération de résumés. À partir de cet exemple, nous pouvons remarquer que l'annotation pose une difficulté au niveau de la densité de l'information à annoter. Seulement pour ces deux phrases, l'annotateur doit être en mesure de relier plusieurs objets temporels entre eux, ce qui devient une tâche laborieuse.

A.2 Représentation graphique d'un document TimeML

Dans le langage TimeML, nous accordons une grande importance aux liens entre les objets temporels, nous représentons un document TimeML par un graphe orienté. Les noeuds du graphe sont les objets temporels et les arcs sont les relations définies dans l'annotation. Ainsi, nous avons trois types de noeuds : `<TIMEX3>`, `<EVENT>`, `<MAKEINSTANCE>` ; et trois types d'arcs : `<TLINK>`, `<SLINK>`, `<ALINK>` ; les `<SIGNAL>` sont considérés comme des étiquettes sur les relations.

Dans la représentation graphique, nous représentons les événements `<EVENT>` et leur réalisation `<MAKEINSTANCE>` par des rectangles que nous distinguons par leur couleur de

contour et d'arrière plan. Nous avons choisi d'utiliser la couleur pour distinguer les événements, car la représentation est réalisée pour être affichée par un écran couleur, malheureusement cette distinction ne peut être faite dans ce document. Par exemple, la phrase

John taught from 1992 through 1995

contient l'<EVENT> *taught* et sa réalisation. L'annotation de cette phrase se trouve dans la figure A.1.

John
<EVENT eid="e1" class="OCCURRENCE" tense="PAST"aspect="NONE"> ***taught*** </EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1" />
<SIGNAL sid="s1"> ***from*** </SIGNAL>
<TIMEX3 tid="t1" type="DATE" value="1992"> ***1992*** </TIMEX3>
<SIGNAL sid="s2"> ***through*** </SIGNAL>
<TIMEX3 tid="t2" type="DATE" value="1995"> ***1995*** </TIMEX3>
<TLINK eventInstanceID="ei1" signalID="s1" relatedToTime="t1"
relType="BEGUN_BY" />
<TLINK eventInstanceID="ei1" signalID="s2" relatedToTime="t2"
relType="ENDED_BY" />

FIG. A.1 – Annotation de la phrase *John taught from 1992 through 1995*

La représentation graphique de cette phrase, générée par l'outil de visualisation de l'annotation est à la figure A.2. Le rectangle (jaune) où est inscrit taught est l'<EVENT> et les deux losanges (bleus) sont des <TIMEX3>. Dans cet exemple, il n'y a pas de relation explicite entre les <TIMEX3>, car ils représentent un temps explicite que nous pouvons comparer et ordonner selon la relation de précédence de gauche à droite. Les liens temporels <TLINK> sont identifiés par un arc en pointillé mauve et étiquetés par le type de lien temporel et le <SIGNAL> entre parenthèses carrées.

Les formes que peuvent prendre les liens temporels de type <SLINK> sont présentes dans la figure A.3. L'arc (rouge) réentrant (NEGATIVE [not]) du noeud *hear* est un <SLINK> introduisant une relation de négation. Lorsque le <SLINK> est affirmatif, l'arc représentant la relation est dessiné en bleu. L'arc réentrant (MODAL [if]) du noeud *leaves* indique une relation de subordination de type modale introduite par le signal *if*. Notons au passage que cet exemple n'est plus valide avec la spécification

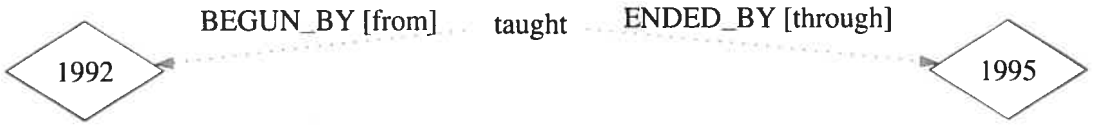

FIG. A.2 – Représentation graphique de la phrase *John taught from 1992 through 1995*, générée par le prototype de représentation visuelle de l'annotation.

actuelle du langage TimeML, il est tout de même pertinent, puisque sa représentation graphique, générée à partir de l'outil de visualisation, nous permet de voir l'influence que la représentation visuelle a eue sur le développement du langage.

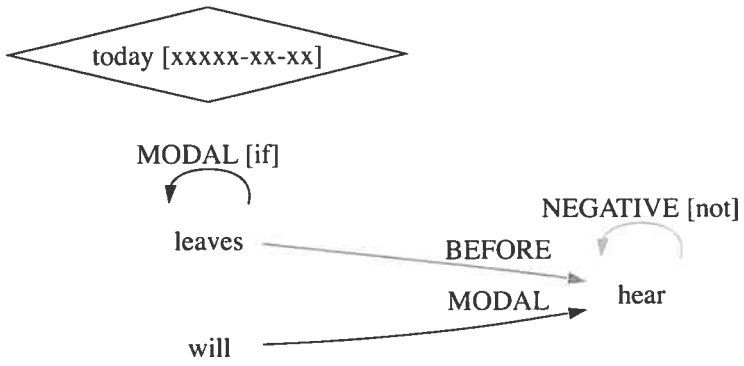

FIG. A.3 – Représentation graphique de la phrase *If Graham leaves today, he will not hear Sabine.*, générée par le prototype de représentation visuelle de l'annotation.

Les exemples précédents n'étaient que de courtes phrases rédigées dans le but de mettre en évidence certaines caractéristiques temporelles. La représentation graphique de l'article de nouvelle suivant[1], annoté manuellement avec la première version du langage est beaucoup plus complète.

> *Qantas will almost double its flights between Australia and India by August in the search for new markets untouched by the crippling Asian financial crisis. This move comes barely a month after Qantas suspended a*

[1] APW19980213.1320 du TimeBank

number of services between Australia, Indonesia, Thailand and Malaysia in the wake of the Asian economic crisis. The airline has also cut all flights to South Korea. Qantas plans daily flights between Sydney and Bombay, up from the current four flights a week, to boost business and tourism ties with India, the airline announced Friday. In a joint statement with Tourism Minister Andrew Thomson, it said two new flights would leave Bombay on Monday and Tuesday nights from March 30, with the third departing each Thursday from August 6. This will add nearly 700 seats a week on the route. Thomson, in India to talk to tourism leaders, said the flights would provide extra support to the growing tourism market. Qantas' India manager Khursheed Lam said the airline was working closely with the Australian Tourist Commission to develop greater awareness of Australia in the Indian market. Qantas will also appoint a Bombay-based public relations consultant.

La représentation graphique de l'article est illustrée par les figures A.4, A.5 et A.6. Cet exemple nous donne un aperçu de l'utilité de la visualisation des documents TimeML, en nous permettant d'identifier rapidement le flot des événements. La représentation graphique de l'annotation permet aux annotateurs de détecter des relations oubliées et d'ordonner temporellement les événements de gauche à droite. Au cours de l'annotation de documents, la visualisation nous a permis de normaliser l'annotation et de corriger des erreurs d'annotation. Nous avons remarqué que les erreurs d'annotation provenaient souvent d'une mauvaise interprétation des spécifications, la représentation graphique nous a servi à clarifier certaines parties ambiguës de la spécification du langage.

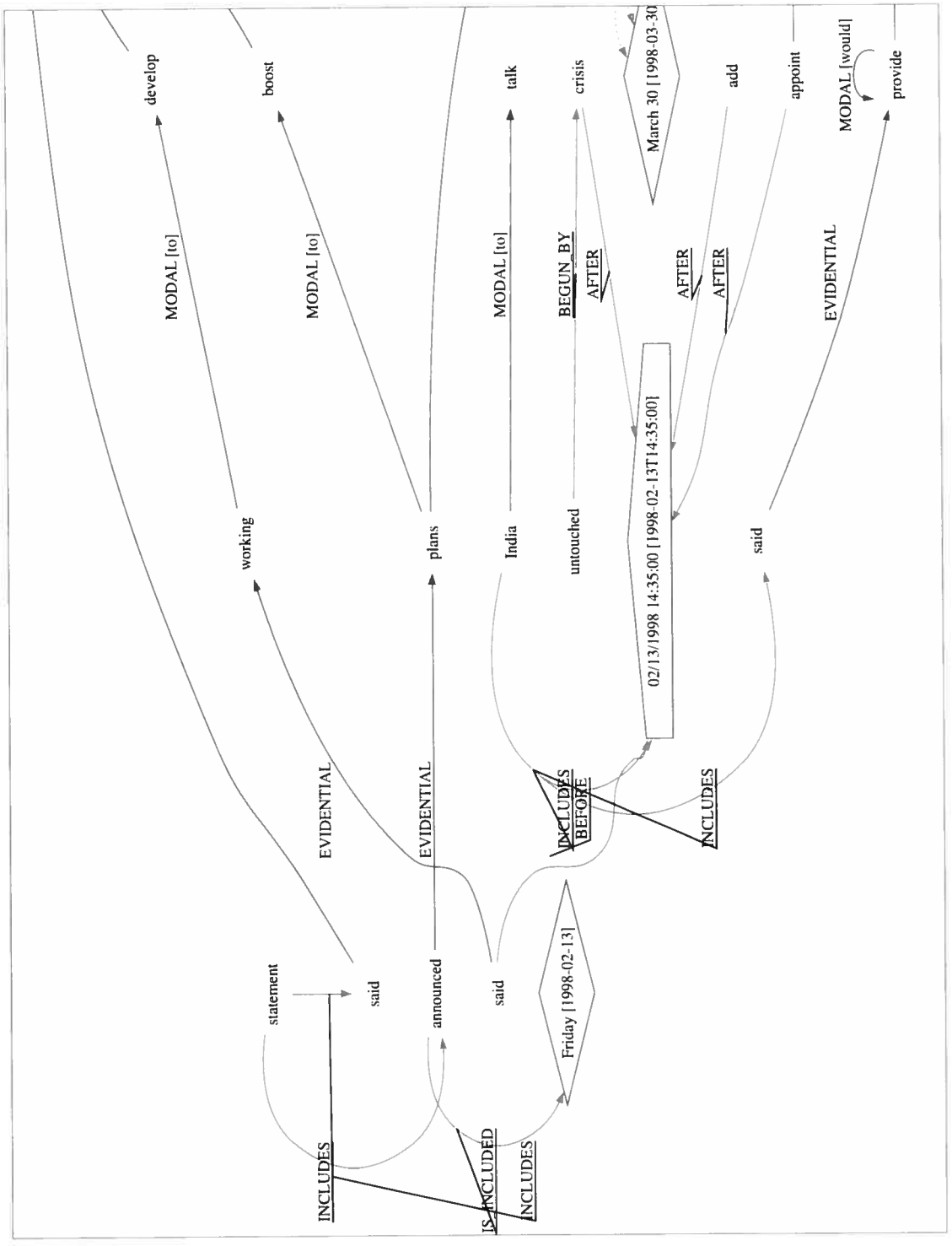

FIG. A.4 – Première partie de la représentation graphique de l'annotation par TimeML du document APW19980213.1320 annoté manuellement provenant du corpus TimeBank

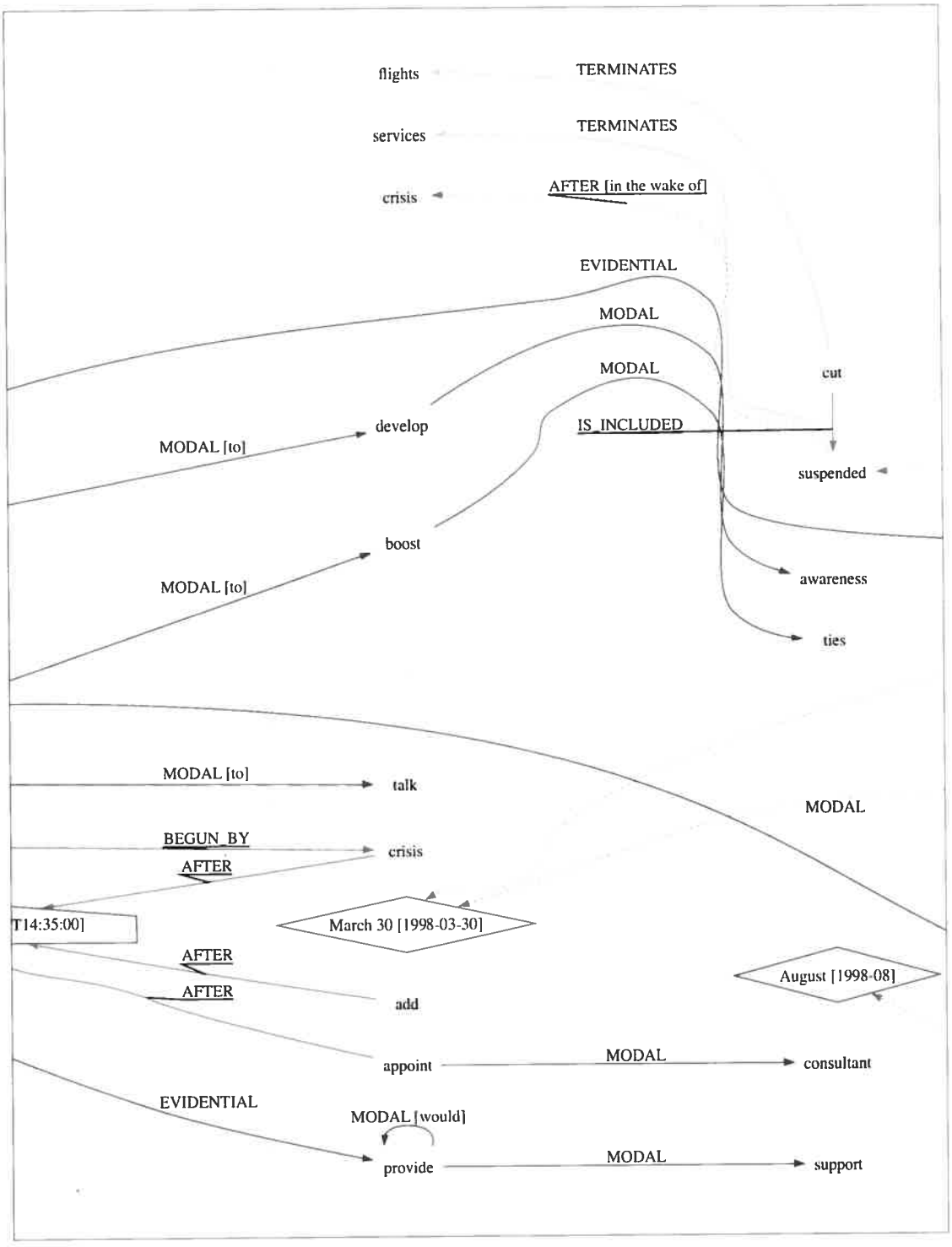

FIG. A.5 – Deuxième partie de la représentation graphique de l'annotation par TimeML du document APW19980213.1320 annoté manuellement provenant du corpus TimeBank

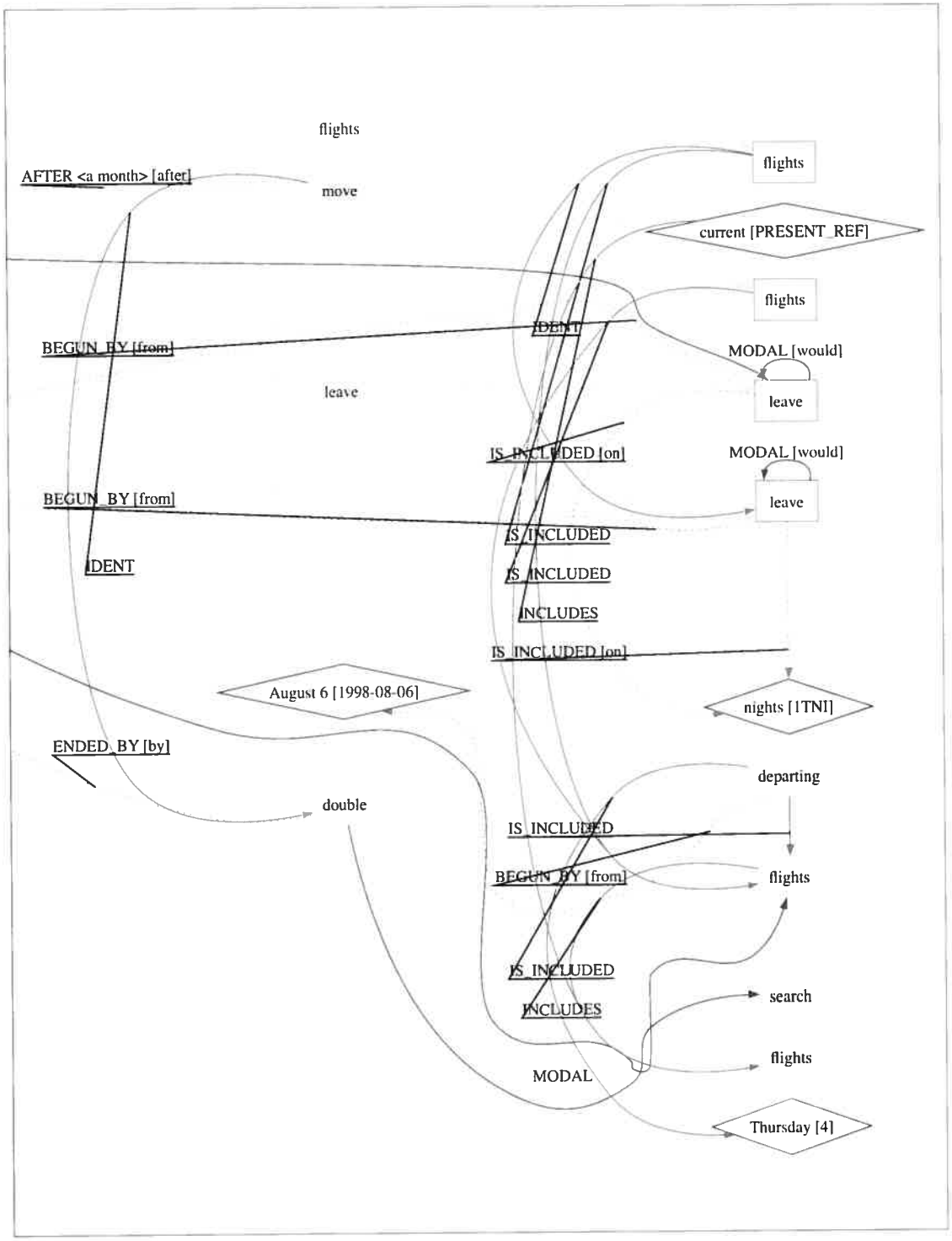

FIG. A.6 – Troisième partie de la représentation graphique de l'annotation par TimeML du document APW19980213.1320 annoté manuellement provenant du corpus TimeBank

A.3 Organisation graphique de l'annotation TimeML

TANGO est une interface graphique qui permet de situer les événements selon leur réalisation dans le temps en ajoutant les liens temporels à l'annotation des événements et en modifiant directement la représentation graphique de l'annotation. TANGO doit être utilisé avec un outil d'annotation traditionnel pour annoter les marqueurs temporels `<TIMEX3>` et les événements `<EVENT>`. Lorsque le document a été annoté avec ces étiquettes, l'ajout des liens temporels se fait à partir de l'interface graphique. L'ajout des `<TLINK>` à l'annotation a une répercussion directe sur la représentation du document, car elles nous permettent d'ordonner les événements. Les `<TLINK>` et `<SLINK>` n'ont pas un effet aussi important sur la représentation graphique, mais en les combinant avec les algorithmes de clôture temporelle [66] leur effet devient visuellement plus apparent.

Nous avons divisé l'interface de TANGO (fig. A.7) en trois aires, la moitié supérieure de la fenêtre contient le texte avec l'annotation des objets temporels, la partie inférieure de la fenêtre est divisée en deux, à gauche nous retrouvons la liste des objets temporels n'ayant pas de liens temporels ; à droite, la représentation graphique de l'annotation.

L'outil d'annotation TANGO permet d'ajouter les liens temporels en choisissant deux objets temporels par une interface pointer-cliquer. La réalisation des événements `<MAKEINSTANCE>` se fait en ajoutant l'événement dans la fenêtre de visualisation, ceci permet de positionner l'événement dans le temps et de rendre possible sa liaison avec un autre événement. Lorsque deux objets temporels sont reliés, une fenêtre surgissante apparaît, nous permettant d'ajouter des attributs à la relation. L'annotateur est aidé de deux outils lors de la réalisation de l'annotation, l'un est la fermeture temporelle des relations, l'autre est la mise en forme automatique du graphe en fonction des relations, de façon similaire à la visualisation statique présentée précédemment. TANGO peut aussi comparer deux annotations temporelles, ceci permet aux annotateurs débutants de comparer leur annotation avec celle de documents déjà annotés et d'évaluer la similarité des annotations.

Nous avons réalisé une description détaillée de l'outil dans le rapport du groupe de travail TANGO [58]. Cet outil est d'une aide précieuse pour les annotateurs, il a

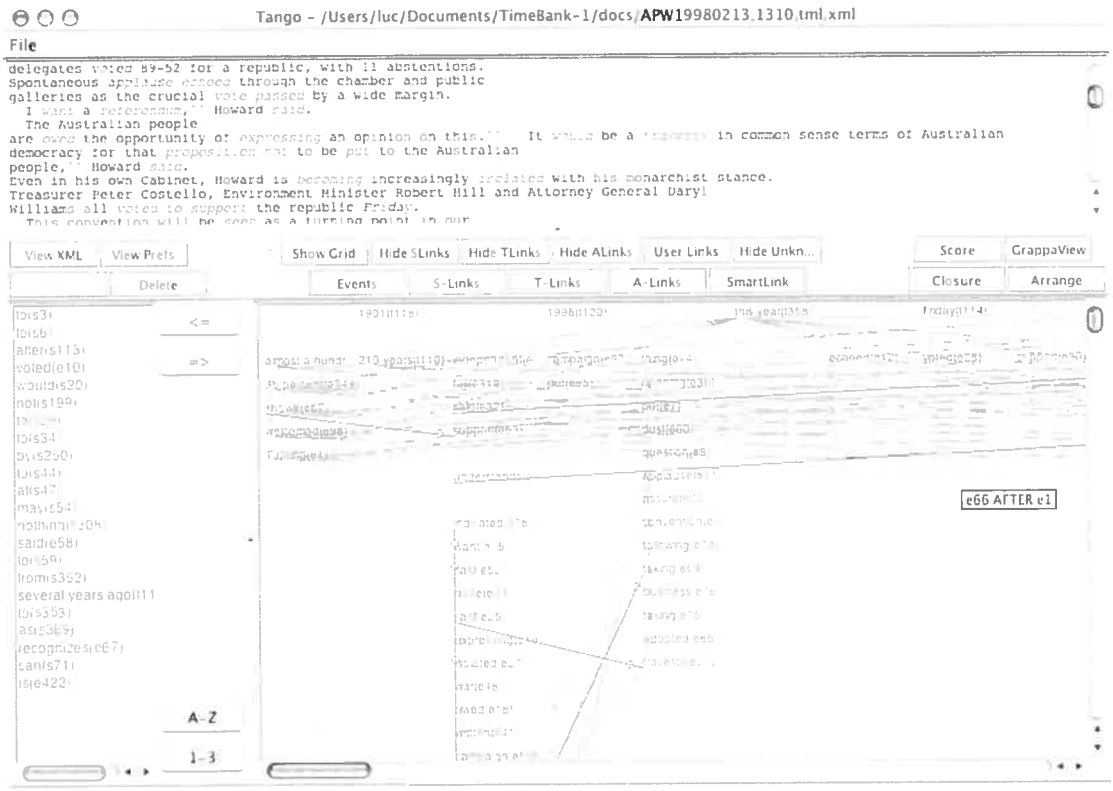

FIG. A.7 – Interface d'annotation graphique TANGO

été utilisé pour l'annotation des liens temporels du corpus TimeBank. Il permet de solutionner plusieurs problèmes dont : les inconsistances dans l'annotation des liens, l'accord entre les annotateurs, la vitesse d'annotation et la validité de l'annotation. TANGO est un élément essentiel pour que le langage TimeML soit utilisé en pratique, il facilite l'annotation des documents en accélérant l'annotation et en augmentant la qualité de l'annotation. La disponibilité d'une grande quantité de documents annotés par TimeML nous permettra de développer des algorithmes pour exploiter les structures temporelles plus efficacement qu'en ce moment.